EE/Ma/CS 126a: Information Theory

Aman Bhargava

October-December 2022

Contents

0.1 Introduction and Course Information

This document offers an overview of EE/Ma/CS 126a at Caltech. They comprise my condensed course notes for the course. No promises are made relating to the correctness or completeness of the course notes. These notes are meant to highlight difficult concepts and explain them simply, not to comprehensively review the entire course.

Course Information

- Professor: Michelle Effros
- Term: 2022 Fall

Chapter 1

Math Review

1.1 Combinatorics & Probability

Binomial Distribution & Coefficient

- **Bernoulli Process**: Repeated trials, each with one binary outcome. The probability of a positive outcome is $p \in [0, 1]$. Each trial is independent.
- **Binomial Distribution**: Let x represent the number of successful trials in a Bernoulli process repeated n times with success probability p. The binomial distribution gives the probability distribution on x:

$$b(x; n, p) = \binom{n}{k} p^x (1-p)^{n-x} \tag{1.1}$$

  Which has $\mu = np$, $\sigma^2 = npq$.

- **Intuition for Binomial Distribution**: The probability of observing a sequence with x positive outcomes and $n - x$ negative outcomes is $p^x(1-p)^{n-x}$. There are $\binom{n}{k}$ different sequences (i.e., permutations) that have x positive cases and n negative cases. Thus the total probability of observing x positive cases is given by Eq 1.1.
- **Binomial Coefficient**:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{1.2}$$

1.2 Logarithm Identities

Entropy calculations and manipulations involve a lot of logarithms. They're not so bad once you get to know them, though:

- **Definition**:
$$a = b^{\log_b a}$$
- **Sum-Product**:
$$\log_c(ab) = \log_c a + \log_c b$$
- **Difference-Quotient**:
$$\log(a/c) = \log a - \log c$$
$$\log \frac{1}{a} = -\log a$$
- **Product-Exponent**:
$$\log_c(a^n) = n \log_c(a)$$
- **Swapping Base**:
$$\log_b(a) = \log_a(b)$$
- **Swapping Exponential**:
$$a^{\log n} = n^{\log a}$$
- **Change of Base**;
$$\log_b(a) = \frac{\log_x(a)}{\log_x(b)}$$

Chapter 2

Entropy Definitions

Chapter 2 of Elements of Information Theory.

2.1 Entropy, Conditional Entropy, Joint Entropy

Entropy Definition (Discrete):

$$H(X) = \sum_{x \in \mathcal{X}} p(x) \log(\frac{1}{p(x)}) \tag{2.1}$$

$$= -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{2.2}$$

$$= \mathbb{E}[\log \frac{1}{p(x)}] \tag{2.3}$$

Theorem 1 *Properties of Entropy*

1. ***Non-negativity:*** $H(X) \geq 0$ – *Reasoning: Entropy is the sum-product of non-negative terms.*
2. ***Change of base:*** $H_b(X) = (\log_b a) H_a(X)$
3. ***Bernoulli entropy:*** $H(X) = -p \log p - q \log q \equiv H(p)$.
    - $H(p)$ *is a concave function of* p*, peaks at* $p = q = 0.5$.

Joint Entropy: Literally just entropy of vector $[X, Y]^\top$.

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \tag{2.4}$$

$$= \mathbb{E}[\log p(x,y)] \tag{2.5}$$

$$\tag{2.6}$$

Conditional Entropy: $H(Y|X)$ is the expected entropy of $p(y|x)$ averaged across all x.

$$H(Y|X) = \sum_{x\in\mathcal{X}} p(x) \underbrace{\sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x)}_{H(Y|X=x)} \tag{2.7}$$

$$= -\mathbb{E}[\log p(Y|X)] \tag{2.8}$$

Entropy can be thought of as the **uncertainty** in the value of a random variable. High entropy corresponds to a high degree of uncertainty. Conditional entropy $H(Y|X)$ can be thought of as the average **remaining uncertainty** in the value of Y after learning the value of X.

Theorem 2 *Chain Rule for Entropy*

$$H(X,Y) = H(X) + H(Y|X) \tag{2.9}$$

$$= H(Y) + H(X|Y) \tag{2.10}$$

It also follows that

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z) \tag{2.11}$$

Proof sketch:

- *Recall that* $H(X) = -\mathbb{E}[\log p(x)]$ *and* $H(Y|X) = -\mathbb{E}[\log p(y|x)]$.
- $\log p(x) + \log p(y|x) = \log(p(x)\cdot p(y|x)) = \log p(x,y)$.
- *The proof follows from there. You can also write out the full sum form of* $H(X,Y)$ *and recover* $H(X), H(Y|X)$ *from there if you're feeling rigorous.*

2.2 Relative Entropy & Mutual Information

Relative Entropy: $D(p\|q)$ gives a *distance* between distributions $p(x)$ and $q(x)$. Also known as KL divergence.

$$D(p\|q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \tag{2.12}$$

$$= \mathbb{E}_{p(x)}[\log\frac{p(x)}{q(x)}] \tag{2.13}$$

$$\tag{2.14}$$

This also corresponds to the **inefficiency** of using q as a replacement for p when generating codes for tokens drawn from $p(x)$.

- **Average code length with correct $p(x)$:** $H(p)$.
- **Average code length with incorrect $q(x)$:** $H(p) + D(p\|q)$.

Theorem 3 *Properties of Relative Entropy*

1. ***Asymmetric:*** *In general,* $D(p\|q) \neq D(q\|p)$.
2. ***Non-negative:*** $D(p\|q) \geq 0$.
3. ***Identity:*** *If* $D(p\|q) = 0$ *then* $p \equiv q$.

We now move onto **mutual information** – a measure of the dependence of two variables. As we will see, it is the **reduction in uncertainty** of X due to knowing Y, on average.

Mutual Information:

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(\frac{p(x,y)}{p(x)p(y)}) \tag{2.15}$$

$$= D\big(p(x,y)\|p(x)p(y)\big) \tag{2.16}$$

$$= \mathbb{E}[\log\frac{p(X,Y)}{p(X)p(Y)}] \tag{2.17}$$

Theorem 4 *Properties of Mutual Information*

- *It is the* ***divergence*** *between* $p(x,y)$ *and* $p(x)p(y)$.
- ***Symmetry:*** $I(X;Y) = I(Y;X)$.

- ***Relation to Entropy:*** *Mutual information is the reduction in uncertainty of each RV expected after discovering the other variable's value.*

$$I(X;Y) = H(X) - H(X|Y) \tag{2.18}$$
$$= H(Y) - H(Y|X) \tag{2.19}$$
$$\tag{2.20}$$

- ***Alternative Entropy Relation:***

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{2.21}$$
$$I(X;X) = H(X) - H(X|X) \tag{2.22}$$
$$= H(X) \tag{2.23}$$

Proof Sketch for (3):

- *Within the definition of* $I(X;Y)$ *there is a term* $\log \frac{p(x,y)}{p(x)p(y)}$.
- *Once you convert the argument of the log into* $p(x|y)/p(x)$*, you can separate out* $H(X) - H(X|Y)$ *using the quotient-difference logarithm rule.*

2.3 Chain Rules: $H(\cdot), I(\cdot;\cdot)$

These chain rules end up being very useful in a lot of proofs. Deeply understanding them is a good idea.

Theorem 5 *Entropy chain rule Let* $X_1, X_2, \ldots, X_n \tilde{p}(x_1, x_2, \ldots, x_n)$*. Then*

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, X_{i-2}, \ldots, X_1) \tag{2.24}$$

Proof: *Repeatedly apply Equation 2.9.*

Intuition of Chain Rule: *It's important to note that the term in the sum is conditioned on elements* X_j *with* $j < i$*. Conditioning always reduces entropy, so it's as though the "additional entropy" from the term must be reduced to account for the previous terms already having been added to the total. Also note that any order can suffice – there is no absolute order in the sum.*