10-704: Information Processing and Learning **Fall 2016**

Lecture 1: August 29

Lecturer: Aarti Singh

Note: *These notes are based on scribed notes from Spring15 offering of this course. LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1.1 About the class

This class focuses on information theory, signal processing, machine learning, and the connections between these fields. Both signal processing and machine learning are about how to extract useful information from signals and data, respectively. The distinction between these fields is vanishing nowadays, however it is still useful to point out some classic differences that still often serve as motivation for the respective fields. Classically signals in signal processing involve a temporal component and are usually transmitted over a channel, while data in machine learning is much more general (but we will see in this course how channels do arise in machine learning problems as well, and help characterize the fundamental limits of achievable error in machine learning problems). A more fundamental difference between signal processing and machine learning is that signals are often *designed* in the former or modeled using an understanding of the system physics. Such physics-driven models can enable inference i.e. answering related questions about the data generating system, whereas in ML, we do not have much control or understanding of the data generating distribution and the models are primarily used to make predictions.

Information theory is a common thread that lays the foundations of signal processing and machine learning, helping us characterize the fundamental limits in these problems - number of bits needed to recover signals accurately and number of samples needed to learn models well. We will start with a brief introduction to information theory, making connections with machine learning (and sometimes signal processing) along the way. Information theory helps us reason about how much information is available in data. Classical information theory studies two main questions:

1. How much information is contained in a signal/data?

   Example 1. *Consider the data compression (source coding) problem.*

   $$\textit{Source} \rightarrow \textit{Compressor} \rightarrow \textit{Decompressor} \rightarrow \textit{Receiver.}$$

   What is the fewest number of bits needed to describe a source message while preserving all the information, in the sense that a receiver can reconstruct the message?

2. How much information can be reliably transmitted through a noisy channel?

   Example 2. *Consider the data transmission (channel coding) problem.*

   $$\textit{Source} \rightarrow \textit{Compressor} \rightarrow \textit{Channel} \rightarrow \textit{Decompressor} \rightarrow \textit{Receiver.}$$

What is the maximum number of bits per channel use that can be reliably sent through a channel, in the sense that the receiver can reconstruct the message with low probability of error? Since the channel is noisy, we will have to add redundancy into the messages. A related question is: how little redundancy do we need so that the reciever can still recover our messages?

Remark 1. *Data compression or source coding can be thought of as a noiseless version of the data transmission/channel coding problem.*

Connection to Machine Learning (ML):

1. Source coding in ML: in ML, the source can be thought of as a model (for example $p(X_1, \ldots, X_n)$) that generates data points $X_1, \ldots, X_n$, and the least number of bits needed to encode this data reflect the complexity of the source or model. Ideas from source coding can be used to pick a descriptive model with low complexity, as in minimum description length (MDL) coding that we will discuss later in the course.

2. Channel coding in ML: The channel specifies a distribution $p(Y|X)$ where X is the input to the channel and Y is the output. For instance, we can view the output $Y = f(X) + \epsilon$ in regression as the output of a noisy channel that takes $f(X)$ as input and adds noise ϵ to it, and the goal is to decode $f(X)$. Classification is similar except that the channel is not additive noise but defines label noise via the transformation $p(Y|X)$ where Y is the label and X is the input. Similarly, in density estimation, X can be a parameter and Y is a sample generated according to $p(Y|X)$ and the goal is to decode X.

We will return to these later in the course and formally show that the answers to the information theory questions also provides answers to the fundamental limits in machine learning problems.

We start by defining some information theoretic quantities more rigorously. Since sources and models are characterized by probability distributions that generate signals and data as random draws, we start by quantifying the information content of a source/model. To do that, we first define the information content of one random draw, and then use it to define the average information content of the distribution from which the random draws are generated.

1.2 Information Content of a Random Outcome

Shannon, the founding father of information theory, defined the information content of a random outcome X to be $\log_2(1/p(X))$ bits. To see why this is a useful definition, let's turn to some examples. We will see that the **Shannon information content** is essentially the minimum number of binary questions (i.e. with two possible answers, say yes/no - corresponding to the base of the logarithm in the definition) needed to figure out the outcome X.

1. We choose an integer from $0 - 63$ uniformly at random. What is the smallest number of yes/no questions needed to identify that integer?

   Intuitive answer: $\log_2(64) = 6$ which can be achieved by doing binary search.

   Information theoretic answer: Note that all outcomes X have probability $p(X) = 1/64$ and hence the Shannon information content of any outcome is $\log_2(64) = \log_2(1/p) = 6$ bits.
   We can also talk about the information content of the outcome of any of the questions, i.e. how much information does each answer provide? Notice that since each answer reduces the search space in half, each outcome of a question has probability $1/2$ and hence information content $\log_2(2) = 1$ bit.

2. In the previous example, the possible answers to each question were equally probable. Here is an experiment that does not lead to equi-probable outcomes. An enemy ship is located somewhere in a 8×8 grid (64 possible locations). We can launch a missile that hits one location. Since the ship can be hidden in any of the 64 possible locations, we expect that we will still gain 6 bits of information when we find the ship. However, each question (firing of a missile) now may not provide the same amount of information since probability of a hit does not necessarily equal the probability of a miss. Notice that we have now *restricted* the binary questions i.e. we are not allowed to ask if the ship is in first 0-31 cells or next 32-63 cells as in previous example. Hence, the number of such *restricted* binary questions needed to find the ship can be more.

   The probability of hitting on first launch is $p_1(h) = 1/64$, so the Shannon Information Content of hitting on first launch is $\log_2(1/p_1(h)) = 6$ bits. Since this was a low probability event, we gained a lot of information (in fact all the infomation we hoped to gain on discovering the ship). However, we will not gain the same amount of information on more probable events. For example:

   The information gained from missing on the first launch is $\log_2(64/63) = 0.0227$ bits.

   The information gained from missing on the first 32 launches is:

$$\begin{aligned}\sum_{i=1}^{32} \log_2(p_i(m)) &= \log_2(\prod_{i=1}^{32} p_i(m)) \\ &= \log_2(\frac{64}{63}\frac{63}{62}\cdots\frac{33}{32}) \\ &= \log_2(2) = 1 \text{ bit}\end{aligned}$$

   This is intuitive, since ruling out 32 locations is equivalent to asking one question in the previous experiment.

   If we hit on the next try, we will gain $\log_2(1/p_{33}(h)) = \log_2(32) = 5$ bits of information. Simple calculation will show that, regardless of how many launches we needed, we gain a total of 6 bits of information whenever we hit the ship.

3. What if the questions are allowed to have more than 2 answers?

   If questions can have k answers, the information content of a random outcome X will be $\log_k(1/p(X))$. Lets understand this with an example.

   Suppose I give you 9 balls and tell you that one ball is heavier. We have a balance and we want to find the heavier ball with the fewest number of weighings. There are now three possible outcomes of an experiment: left side heavier, right side heavier, or equal weight. Since any of the 9 balls can be heavier, the probability of each outcome is 1/9 and the Shannon information content is then $\log_3$(number of balls) $= \log_3(9) = 2$.

   This is indeed the same as the minimum number of experiments (weighings) needed. The strategy is to split the balls into three groups of three and weigh one group against another. If these two are equal, you split the last group into three and repeat. If the first group is heavier, you split it and repeat and vice versa for the second group.

   Note that to gain a lot of information, you want to design experiments so that each outcome is equally probable.

   Suppose now that the odd ball is either heavier or lighter, but I don't tell you which. How many weighings do we need? There are 18 possible outcomes so information theory tells us $\log_3(18)$ experiments would be necessary, and this number is between 2 and 3. It says that the minimum number of experiments is more than 2 but no more than 3. This brings up an interesting point: Note that we may need more (3) than information theoretic limit ($\log_3(18)$); and we will indeed see that information theoretic bounds are often not achievable but they still provide us fundamental limits.

1.3 Information Content of a Random Variable

A random variable is an assignment of probability to outcomes of a random experiment. We then define the information content of a random variable as the expected Shannon Information Content of the outcomes. This is referred to as the **Entropy** of a random variable.

Definition 2. *The **entropy** of a random variable X with probability distribution $p(x)$ is:*

$$H(X) = \sum_{x \in \mathcal{X}} p(x) \log_2(1/p(x)) = -\mathbb{E}_{X \sim p}[\log_2 p(X)], \tag{1.1}$$

where $\mathcal{X}$ is the set of all possible values of X. We often write $H(X) = H(p)$ since entropy is a property of the distribution.

We can also think of it as a measure of uncertainty about the random variable. To see this, lets compute the entropy of some simple distributions:

- If $X \sim \text{Uniform}(\mathcal{X})$, then $H(X) = \sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} \log_2(|\mathcal{X}|) = \log_2(|\mathcal{X}|)$. In this case, the more the number of outcomes, the less certain we are and hence the entropy (uncertainty) increases increases with the number of outcomes.
- If $X \sim \text{Bernoulli}(p)$ then $H(X) = H(p) = -p \log_2 p - (1-p)\log_2(1-p)$. In this case, the outcomes are always two but if $p = 0$ or $p = 1$ the outcome is highly certain - always 0 (Tail) or 1 (Head), respectively. In this case, the entropy is zero. On the other hand, if $p = 1/2$ the outcome is highly uncertain and entropy is higher (1 bit - we gain one bit of information when we know the outcome).