10-704: Information Processing and Learning **Spring 2012**

Lecture 1: Introduction, Entropy and ML estimation

Lecturer: Aarti Singh *Scribes: Min Xu*

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

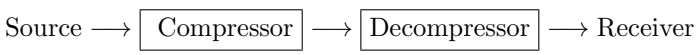
This class focuses on information theory, signal processing, machine learning, and the connections between these fields.

- *signals* can be audio, speech, music. *data* can be images, files. They overlap a lot of times.
- Both signal processing and machine learning are about how to extract useful information from signals/data.
- *signals* as used in the EE community can be different from *data* in that (1) they often have temporal aspect, (2) they are often designed and (3) they are often transmitted through a medium (known as a *channel*).

Information theory studies 2 main questions:

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

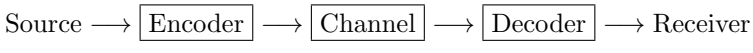
   Example: Consider the data compression (source coding) problem.

   What is the fewest number of bits needed to describe the output of a source (also called message) while preserving all the information, in the sense that a receiver can reconstruct the message from the bits with arbitrarily low probability of error?

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

   Example: Consider the data transmission (channel coding) problem.

   What is the maximum number of bits per channel use that can be reliably sent through a noisy channel, in the sense that the receiver can reconstruct the source message with arbitrarily low probability of error?

Remark: The data compression or source coding problem can be thought of as a noiseless version of the data tranmission or channel coding problem.

Connection to Machine Learning:

1. *Source Coding in ML*: In ML, the source is essentially a model (e.g. $p(X_1, ..., X_n)$) that generates data points $X_1, ..., X_n$, and the least number of bits needed to encode these data reflect the complexity of the source or model. Thus, source coding can be used to pick a descriptive model with the least complexity. This is the principle of Occam's Razor.
2. *Channel Coding in ML*: The channel specifies a distribution $p(y \mid x)$ where x is the input to the channel and y is the output. For instance, we can view the output $y_i = m(x_i) + \epsilon$ in regression as the output of a noisy channel that takes $m(x_i)$ as input. Similarly, in density estimation, x can be a parameter and y is a sample generated according to $p(y \mid x)$.

We will formalize these notions later in this course.

1.2 Information Content of Outcomes of Random Experiments

We will usually specify information content in bits, where a bit is defined to be of value either 0 or 1. We can think of it as the output of a yes/no question, and the information content can be specified as the minimum number of yes/no questions needed to lean the outcome of a random experiment.

1. We have an integer chosen randomly from 0 to 63. What is the smallest number of yes/no questions needed to identify that integer? Answer: $\log_2(64) = 6$ bits.

   Note: all integers from 0 to 63 have equal probability $p = \frac{1}{64}$ of being the correct integer. Thus, $\log_2(64) = \log_2(\frac{1}{p})$. This is the **Shannon Information Content**.

2. Now lets consider an experiment where questions that do not lead to equi-probable outcomes.

   An enemy ship is somewhere in an 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.

   The probability of hitting on first launch is $p_1(h) = \frac{1}{64}$, so the Shannon Information Content of hitting on first launch is $\log_2(\frac{1}{p_1(h)}) = 6$ 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 probably events. For example:

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

   The information gained from missing on the first 32 launches is

$$\begin{aligned}\log_2(p_1(m)) + \log_2(p_2(m)) + ... \log_2(p_{32}(m)) &= \log_2(\prod_{i=1}^{32} p_i(m)) \\ &= \log_2(\frac{64}{63}\frac{63}{62}...\frac{33}{32}) \\ &= \log_2(2) = 1 \text{ bit}\end{aligned}$$

   Which is what we intuitively expect, since ruling out 32 locations is equivalent to asking 1 question in the previous experiment 1.

If we hit on the next try, we will gain $\log_2(\frac{1}{p_3 3(h)}) = \log_2(32) = 5$ 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?

   Suppose we have a number of identically looking balls one of which is either heavier or lighter. We have a balance and we want to find the odd ball with fewest number of weighings. There are now three possible outputs of an experiment: left lighter, left heavier, equal weight. The minimum number of experiments we need is then $\log_3$(number of balls). Note that we may need more; information bounds are often not achievable.

1.3 Information Content of Random Variables

A random variable is simply an assignment of probability to outcomes of a random experiment. We then define the information content of a random variable as just the average Shannon Information Content. We can also think of it as a measure of the uncertainty of the random variable.

Definition 1.1 *The **entropy** of a random variable X with probability distribution $p(x)$ is*

$$H(X) = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{1}{p(x)} = -\mathbb{E}_p[\log p(x)]$$

Where $\mathcal{X}$ is set of all possible values of the random variable X. We often also write $H(X)$ as $H(p)$ since entropy is a property of the distribution.

Example
For a Bernoulli random variable with distribution $Ber(p)$. The entropy $H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$

Definition 1.2 *Suppose we have random variables X, Y, then the **Joint Entropy** between them is*

$$H(X, Y) = -\sum_{x,y} p(x, y) \log_2 p(x, y)$$

This is a measure of the total uncertainty of X, Y.

Note: If X, Y are independent, then it is easy to show that $H(X, Y) = H(X) + H(Y)$. If X, Y are dependent, then $H(X, Y) < H(X) + H(Y)$ in general.

Definition 1.3 *Suppose we have random variables X, Y, then the **Conditional Entropy** of Y conditioned on X is defined as*

$$\begin{aligned} H(Y \mid X) &= \sum_x p(x) H(Y \mid X = x) \\ &= -\sum_x p(x) \sum_y p(y \mid x) \log_2 p(y \mid x) \\ &= -\sum_x \sum_y p(x, y) \log_2 p(y \mid x) \\ &= -\mathbb{E}_{X,Y}[\log \frac{1}{p(y \mid x)}] \end{aligned}$$

The conditional entropy is the average uncertainty in Y after we had observed X.

Theorem 1.4 *(Chain Rule)*

$$H(X,Y) = H(Y \mid X) + H(X)$$

The proof of the chain rule will be covered in recitation on in the homework.

Definition 1.5 *Given two distributions p, q for a random variable X. The **Relative Entropy** between p and q is defined as*

$$\begin{aligned} D(p \,||\, q) &= \mathbb{E}_p[\log \frac{1}{q}] - \mathbb{E}_p[\log \frac{1}{p}] \\ &= \mathbb{E}_p[\log \frac{p}{q}] \\ &= \sum_x p(x) \log \frac{p(x)}{q(x)} \end{aligned}$$

The relative entropy is also known as **Information divergence** or **KL (Kullback-Leibler)** divergence. The relative entropy is the cost incurred if we used distribution q to encode X when the true underlying distribution is p. We will make this intuition more precise later in the course. We note that relative entropy has some important properties:

- $D(p \,||\, q) \geq 0$, and equals 0 iff $p = q$
- Relative entropy is often not symmetric, i.e., $D(p \,||\, q) \neq D(q \,||\, p)$

Definition 1.6 *Let X, Y be two random variables. The **Mutual Information** between X and Y is the following:*

$$I(X,Y) = D(p(x,y) \,||\, p(x)p(y))$$

where $p(x,y)$ is the actual joint distribution of X, Y and $p(x)$ and $p(y)$ are the corresponding marginal distributions. Thus, $p(x)p(y)$ denotes the joint distribution that would result if X, Y were independent.

Note: we can give a preview of the answers of the two fundamental questions of information theory stated before.

1. How much can we compress data? If the data is generated as a random variable X, then the answer is $H(X)$.
2. How much data can we transmit over a noisy channel? If X is the input and Y is the output, then the answer is $\max_{p(X)} I(X,Y)$.

We will prove these later.

1.4 Connection to Maximum Likelihood Estimation

Suppose $X = (X_1, ..., X_n)$ are data generated from a distribution $p(X)$. In maximum likelihood estimation, we want to find a distribution q in some family of distributions $\mathcal{Q}$ such that the likelihood $q(X)$ is maximized:

$$\max_{q \in \mathcal{Q}} q(X) = \min_{q \in \mathcal{Q}} -\log q(X)$$

In machine learning, we often define a loss function. In this case, the loss function is the negative log loss: $\text{loss}(q, X) = -\log q(X)$. The expected value of this loss function is the risk: $Risk(q) = \mathbb{E}_p[\log \frac{1}{q(x)}]$. We want to find a distribution q that minimizes the risk. However, notice that minimizing the risk with respect to a distribution q is exactly minimizing the relative entropy between p and q. This is because

$$Risk(q) = \mathbb{E}_p[\log \frac{1}{q(x)}] = \mathbb{E}_p[\log \frac{p(x)}{q(x)}] + \mathbb{E}_p[\log \frac{1}{p(x)}] = D(p||q) + Risk(p)$$

Because we know that relative entropy is always non-negative, we know here that the risk is minimized by setting q equal to p. Thus the minimum risk $R^* = Risk(p) = H(p)$, the entropy of distribution p. The excess risk, $Risk(q) - R^*$ is precisely the relative entropy between p and q.