Entropy And The Mind

Contents

1 Probability Primer

1.1 Measure theory

Measure theory tells us what is measurable and what is not measurable. The need for such theory is non-obvious. For why do we need to know what is measurable and what is not. An example hopefully should show the reason.

Let's take a ball and cut it into a infinite but countable number of pieces. Using translation and rotation we can arrange the pieces in a way which will create two balls identical to the first ball. This can be done using the most minimal assumptions. Now we can say two things about this case as to assume

why it is false. First we can say that the assumptions we had were wrong or second, we can say that the pieces that we cut up are non-measurable. The first options is unlikely since we choose the most minimal set of assumptions yet the second seems a likely option. This Banach-Tarski Paradox is a example for things that we might define as non-measurable. To define what is measurable and what is non-measurable we have to first take a look into σ-algebra.

1.1.1 Sigma-Algebra

Definition 1 *Given a set Ω, a power set on Ω is the set that contains the empty set, the set itself, and every combination on the set.*

For example we can look into the set $\Omega \in \{0,1\}$.

$$\Omega \in \{0,1\} \implies 2^{\Omega} = \{\emptyset, \{0\}, \{1\}, \{0,1\}\}.$$

Definition 2 *Given a set Ω, a σ-algebra on Ω is a collection $A \subset 2^{\Omega}$ s.t. A is non-empty and A is:*

1. *Closed under complements* [1] $(E \in A \implies E^c \in A)$
2. *Closed under countable unions* $(E_1, E_2, \cdots \in A \implies \bigcup_{i=0}^{\infty} E_i \in A)$

Remark 1 .

1. $\Omega \in A$ *since* $E \in A \implies E^c \in A \implies E \cup E^c \in A \implies \Omega \in A$
2. $\emptyset \in A$ *since* $\Omega \in A \implies \Omega^c \in A \implies \emptyset \in A$
3. *A is close under countable intersections. Suppose* $E_1, E_2, \cdots \in A$.
   $\cap_{i=0}^{\infty} E_i = \cap (E_i^c)^c = (\cup E_i^c)^c \in A$

Definition 3 *Given $C \subset 2^{\Omega}$, the σ-algebra generated C, written $\sigma(C)$, is the "smallest" σ-algebra containing C.*

$$\text{That is, } A \text{ is a } \sigma - \text{algebra, } \sigma(C) = \bigcup_{A \supset C} A.$$

Remark 2 $\sigma(c)$ *always exists because:*

1. 2^{Ω} *is a σ-algebra*
2. *any intersection of σ-algebras is a σ-algebra*

[1] A complement of E is made out of all elements which are missing in E but are still in the universe(Ω) of E. e.g. $A = \Omega = \{1,2,3,4\}$, $B \subset A = \{1,2\} \implies B^c = \{3,4\}$

1.1.2 Measure

To define what is a probability measure we first have to define what is a measure.

Definition 4 *A measure μ on Ω with σ-algebra A is a function $\mu : A \to [0, \infty]$ s.t.*

1. $\mu(\emptyset) = 0$
2. $\mu(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mu(E_i)$ *for any $E_1, E_2, \cdots \in A$ pairwise disjoint sets*[2] *In other words, "countable additivity"*

Kolmogorov's axioms are a way of formalizing probability theory. They help define what is a probability measure and some of it's properties.

Definition 5 *A probability measure is a measure P s.t. $P(\Omega) = 1$.*

Now that we know what is a measure, let's take a look into examples of measures.

1. (Finite set) $\Omega = \{1, 2, \ldots, n\}$, $A = 2^{\Omega}$, $P(\{k\}) = P(k) = \frac{1}{n}$, $\forall k \in A \geq 0$. (Uniform Distribution) $P(\{1, 2, 4\}) = P(\{1\} \cup \{2\} \cup \{4\}) = P(1) + P(2) + P(4)$
2. (Countably infinite) $\Omega = \{1, \ldots, n\}$, $P(k) =$ probability that a coin flip will be heads after k coin flips. $P(k) = \alpha(1-\alpha)^{k-1}$. For a fair coin $\alpha = \frac{1}{2}$, $P(k) = \frac{1}{2}(1 - \frac{1}{2})^{k-1} = (\frac{1}{2})^k$, $\forall k \geq 1$. We can show P is a probability measure because $\Omega = \{P(1), P(2), \ldots, P(n)\} \implies \Omega = \{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots\} \implies$ geometric series $\implies$ $P(\Omega) = \sum \Omega = \frac{a}{1-q} = \frac{0.5}{1-0.5} = 1$. (Geometric Distribution)
3. (Uncountable) $\Omega[0, \infty)$, $A = B([0, \infty))$, $P([0, x)) = 1 - e^x$ $\forall x > 0$ (Exponential distribution). Note: $P(\{x\}) = 0$

Theorem 1 *(Basic properties of measures). Let (Ω, A, μ) be a measure space.*

1. *Monotinicity: If $E, F \in A$ and $E \subset F$ then $\mu(E) \leq \mu(F)$.*
2. *Subadditivity: If $E_1, E_2, \cdots \in A$ then $\mu(\bigcup_{i=1}^{\infty} E_i) \leq \sum_{i=1}^{\infty} \mu E_i$*

1.2 Probability

Terminology: event = measureable set = set in A. sample space = Ω.

1.2.1 Conditional Probability

Definition 6 *if $P(B) > 0$ then $P(A|B) = \frac{P(A \cap B)}{P(B)}$.*

This defines a conditional probability. Which essentially says, the probability of A given B is the probability of getting a A inside of the space formed by B (as we know that B has occurred).

[2] Pairwise disjoint sets are sets which don't have an intersection. $A = \{1, 2, 3, 4, 5\}$, $B \subset A = \{1, 2\}$, $C \subset A = \{4\}$. In this case, set B and set C are pairwise disjoint, they don't have any common elements meaning $B \cap C = \emptyset$.

1.2.2 Independence

Definition 7 *Event A, B are independent if $P(A \cap B) = P(A)P(B)$*

Definition 8 *$A_1, A_2, \ldots, A_n$ are mutually independent if any $S \subset \{1, \ldots, n\}$*

$$P(\bigcap_{i \in S} A_i) = \prod_{i \in S} P(A_i).$$

It is important to note that mutual independence $\implies$ pairwise independence but the opposite is untrue.

Definition 9 *A, B are conditionally independent given C if*

$$P(A \cap B|C) = P(A|C)P(B|C) \text{ if } P(C) > 0.$$

Now having defined 3 independence types, let us look into the following proposition: Suppose $P(B) > 0$. Then A, B are independent $\iff P(A|B) = P(A)$.
Proof: $P(A \cap B) = P(A)P(B) \implies P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$.

1.2.3 Bayes equation

Bayes equation tells us how to update our expectations given new information. The known formula hides within the intuition behind it. For that reason aside from the main formula a more thorough look is required.

Theorem 2 *Bayes' rule*

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}, \text{ if } P(B) \neq 0.$$

To gain the intuition I will first derive it, and then explain further the intuition behind these equations.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ if } P(B) \neq 0$$
$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \text{ if } P(A) \neq 0$$

From these if we substitute in both cases $P(A \cap B)$ we get the Bay's theorem.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}, \text{ if } P(B) \neq 0.$$

From this equation it is hard to see why this is the case. For the intuition behind the formula we would have to turn to a more basic look for what is happening. Let's take for example two water fountains in two colors red and blue. The red fountain is on the right separated from the blue fountain on the left by a barrier. The water from the fountains then flows down, with some water meeting at the

end and some going to the side. 80% of the water is blue and 20% is red. The probability of the red water to go to the side is 20%, the probability of the blue water to go to the side is 60%. The question is then, out of the purple water what is the probability of red to to into the purple. Stated differently:

$$P(\text{red}|\text{purple} = 1) =?$$

To answer let us look at the prior odds. The prior odds are 20% red, 80% blue. Then we get to know that 20% of all the red water is going to the side so, 80% from the red water is going down. We also know that 60% of the blue water is going to the side and so 40% of the blue water reaches down.

$$P(\text{red}) = 0.2, P(\text{down}|\text{red} = 1) = 0.8 \tag{1}$$

$$P(\text{blue}) = 0.8, P(\text{down}|\text{blue} = 1) = 0.4 \tag{2}$$

$$P(\text{down}|\text{red}) = \frac{P(\text{down} \cap \text{red})}{P(\text{down})}, P(\text{down}) \neq 0 \tag{3}$$

$$P(\text{down}) = P(\text{red} \cap \text{down}) + P(\text{blue} \cap \text{down}) \tag{4}$$

$$P(\text{down}|\text{red}) = \frac{P(\text{red}) * P(\text{down}|\text{red} = 1)}{P(\text{red}) * P(\text{down}|\text{red} = 1) + P(\text{blue}) * P(\text{down}|\text{blue} = 1)} \tag{5}$$

The answer above is the hard way of answering this problem. It relies solely on the equations that make up the problem without intuition. Now I will show the intuition behind the equations. To do so, we will have to think in terms of odds instead of probabilities.

With out first example, the ratio or odds between the red and blue water at the beginning was $\frac{20\%}{80\%} = \frac{1}{4}$ or $1 : 4$. The ratio between the red water that goes down and blue water the goes down is $\frac{80\%}{40\%} = \frac{2}{1}$ or $2 : 1$. If we then multiply these odds we get $1 : 4 \cdot 2 : 1 = 1 : 2$. This means that the ratio of red water to blue water down at the end is $1 : 2$. From this we can easily calculate the probabilities we wanted. The probability of the water to be red if it is taken from the purple water is $\frac{1}{1+2} = \frac{1}{3}$. A more formal description of what has happened: The function P returns the probability of a given event happening. It should be noted that $P(event) \in \mathbb{R}$, $0 \leq P(event) \leq 1$. Given hypothesis H and event e, Baye tells us that:

$$\frac{P(H_x|e_y)}{P(\neg H_x|e_y)} = \frac{P(H_x)}{P(H_y)} \cdot \frac{P(e_y|H_x)}{P(e_y|\neg H_x)}$$

The hypothesis H is in-fact the state you hypothesis is the case before a event. Event e is the event which tells us new information we use to update our previous beliefs on the hypothesis. Essentially what we are seeing is the following:

$$\text{posterior odds} = \text{prior odds} \cdot \text{likelihood ratio}$$

The prior odds in which we say what we currently think about the hypothesis. The likelihood ratio which tells us the strength of the evidence to allow for updating of our belief. And the posterior odds, which tells us what are the resulting odds after we have updated our belief given the event.

The resources used:
https://arbital.com/p/bays_rule_guide

1.2.4 Chain Rule

Theorem 3 *if* $A_1, \ldots, A_n$ *and* $P(A_1 \cap \cdots \cap A_{n-1}) > 0$,
$(*)$ $P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \ldots P(A_n \cap \cdots \cap A_{n-1})$

Proof 1 *By induction. If* $n = 2$, *then* $(*)_n$ *holds. Suppose* $(*)_{n-1}$. *Let* $B = A_1 \cap \cdots \cap A_{n-1}$. $P(B) = P(A_1)P(A_2|A_1) \ldots P(A_{n-1}|A_1 \cap \cdots \cap A_{n-1})$
$P(B \cap A_n) = P(A_n|B)P(B) =$ *(the above)* $P(A_n|A_1 \cap \cdots \cap A_{n-1})$

1.2.5 Partition

Definition 10 *A partition of* Ω *is a finite or countable collection* $\{B_i\} \subset 2^\Omega$ *s.t.*

1. $\bigcup_i B_i = \Omega$
2. $B_i \cup B_j = \emptyset$ $(i \neq j)$

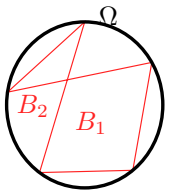


Figure 1: partition

Theorem 4 *Partition rule:* $P(A) = \sum_{i=0} P(A \cap B_i)$ *for any partition* $\{B_i\}$ *of* Ω.

Proof 2

$$A = A \cap \Omega = A \cap (\cup_i(B_i)) = \cup_i(A \cap B_i).$$
$$P(A) = P(\cup_i(A \cap B_i)) = \sum_i P(A \cap B_i).$$

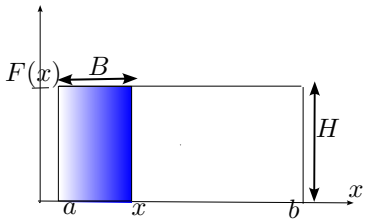
1.3 Random Variables

Definition 11 *Given* (Ω, A, P) *a random variable is a function* $X : \Omega \to \mathbb{R}$ *s.t.* $\{\omega \in \Omega : X(\omega) \leq x\} \in A$, $\forall x \in \mathbb{R}$.

A random variable is a function for quantifying the outcome of a random process. e.g. let us take a fair coin and flip it. We want to measure the amount of times the coin was heads. The random variable in this case maps the set of heads and tails of the coins into a sum of the amount of heads in that set $X : \Omega \to \mathbb{R}$. The benefit here is that random variables allow us to work on measurable sets.

There is also notation that is common to find yet is confusing if not known. Some of the notation is that X, Y are random variables while x, y are the values they take on. Some short hands that people use are: $\{X \leq x\} = \{\omega \in \Omega : X(\omega) \leq x\}$ and also $P(X \leq x) = P(\{X \leq x\})$.

1.3.1 PDF and CDF



The function in the figure is $f(x) = \frac{1}{b-a}$ essentially a uniform distribution. This function represents the "Probability Distribution Function" and in this case is a uniform distribution. The are under the PDF is essentially the CDF. In other words, the CDF of $f(x)$ is

$$\text{CDF} = \int_a^x f(x)dx = H \cdot B = \frac{1}{b-a}(x-a) = \frac{x-a}{b-a}.$$

Definition 12 *The CDF of a random variable X is the function $F : \mathbb{R} \to [0, 1]$ s.t. $F(x) = P(X \leq x)$*

This definition says the following: $CDF = P(X \leq x)$. e.g. lets take the CDF of 1.3.1 which we know to be $\frac{x-a}{b-a}$. The definition says that the CDF $F(x)$ can be rewritten as $P(X \leq x)$ which in our example is the same as $\frac{x-a}{b-a}$.

Definition 13 *The distribution of X is the probability measure P^X on $\mathbb{R}$ s.t. $P^X(A) = P(X \in A)$. $\forall A \in B(\mathbb{R})$.*

The two definitions above have a connection between them. P^X is the probability measure induced F (CDF of X). Proof to the claim: $Q((-\infty, x]) = F(x) = P(X \leq x) = P(X \in (-\infty, x]) = P^X((-\infty, x]) \implies Q = P^X$

1.3.2 Discrete and continues Random Variables

The types of random variables people usually encounter are discrete random variables and continues random variables. These are the two main types of random variables.

Definition 14 *A random variable X is discrete if $X(\Omega)$ is countable.*

In this case $X(\Omega) = \{X(\omega) : \omega \in \Omega\}$. And the fact that it is countable just means $X(\Omega)) = \{x_1, x_2, \ldots\}$

Definition 15 *A random variable X has a density f if $F(x) = \int_{-\infty}^{x} f(a)du$ $\forall x \in \mathbb{R}$. for some integrable $f : \mathbb{R} \to [0, \infty]$.*

Definition 16 *Let $Q = P^X$ and let $J = \{x \in \mathbb{R} : Q(x) > 0\}$.*

$$\left.\begin{array}{l} Q_d(A) = Q(A \cup J) \\ Q_c(A) = Q(A) - Q(A \cup J) \end{array}\right\} \implies Q = Q_d + Q_c$$

Here we have a definition of CDF made up of Discrete and Continues parts. Where, Q_d is the discrete part and Q_c is the continues part.

1.3.3 Discrete random variable

Definition 17 *The PMF (probability mass function) of a discrete random variable X is the function $p(x) : \mathbb{R} \to [0, 1]$ s.t. $p(x) = P(X = x)$.*

In essence the PMF tells us the probability of some value of a random variable event occurring. I will now provide some examples of distributions and their PMF's.

1. $X \sim$ Bernaulli(α), $\alpha \in [0, 1]$ $p(1) = \alpha$, $p(0) = 1 - \alpha$.
2. $X \sim$ Binomial(α), $\alpha \in [0, 1]$ $p(k) = (n, k)\alpha^k(1 - \alpha)^{n-k}$
   where $k \in \{0, 1, \ldots, n\}$. (n, k) $= \frac{n!}{k!(n-k)!}$
3. $X \sim$ Geometric(α), $\alpha \in [0, 1]$ $p(k) = \alpha(1 - \alpha)^{k-1}$ where $k \in \{1, 2, 3, \ldots\}$
4. $X \sim$ Poisson(λ), $\lambda \geq 0$. $p(k) = e^{-\lambda}\frac{\lambda^k}{k!}$ where $k \in \{0, 1, 2, \ldots\}$

Notation: some notes about the notation.
$X \sim p$ this means that X is distributed according to the PMD p. (Ω, A, P)
$X \sim F$ if F is a CDF this means that X has the CDF F
$X \sim Q$ when Q is the distribution of X written as P^X. X has that distribution.

1.3.4 Random variables with densities

A random variable with a density is a random variable X that satisfies the following: $F(X) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$.
Some more useful notation:

1. we call f the probability density function (PDF) of x, and write $X \sim f$.
2. "indicator function of A" $I_A(x) = \begin{cases} 0 & \text{if } x \in A \\ 1 & \text{otherwise} \end{cases}$

Examples of discrete random variable densities:

1. $X \sim \text{Uniform}(a,b)$, $a < b : f(x) = \frac{1}{b-a}$ $x \in [a,b]$ and $f(x) = 0$ otherwise. It looks like 1.3.1
2. $X \sim \text{Exponential}(\lambda)$, $\lambda > 0 : f(x) = \lambda e^{-\lambda x}$, $x \geq 0$. This distribution has the property that it is "memoryless".
3. $X \sim \text{Beta}(\alpha, \beta)$, $\alpha > 0, \beta > 0 : f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$, $X \in [0,1]$
4. $X \sim \text{Normal}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, $x \in \mathbb{R}$. this is also called the "Gaussian" distribution.

1.3.5 Expectation

Definition 18 *The expectation of a discrete random variable X with PMF p is*

$$E(X) = \sum_{x \in X(\Omega)} x \cdot p(x) \text{ when this sum is "well-defined".}$$

Otherwise, expectation does not exist.

e.g. $X \sim \text{Bernauli}(\alpha)$, $X(\Omega) = \{0,1\}$. $E(X) = 0p(0) + 1p(1) = \alpha$

Definition 19 *The expectation of random variable X with density f is*

$$E(X) = \int_{-\infty}^{\infty} x(fx)\ dx.$$

when this integral is "well defined". Otherwise, the expectation does not exist.

e.g. $X \sim \text{Uniform}(a,b)$ has a $p(x) = \frac{1}{b-a}$.

$$\begin{aligned} E(X) &= \int_a^b x \cdot p(x) dx = \frac{1}{b-a} \int_a^b x\ dx \\ &= \frac{1}{b-a} \frac{x^2}{2} = \frac{x^2}{2(b-a)} \implies \frac{b^2-a^2}{2(b-a)} \\ &= \frac{b+a}{2}. \end{aligned}$$

Here we can see that the average value to expect in a uniform distribution is the simple average over the range of values the function takes. The general intuition behind the general formula is that it gives the average value you can expect from the function in a more general formal way that can be defined for more complex functions.

If we back trace a bit and remind ourselves the definition of the expectation, We might notice that it is unclear what is the meaning of a "well-defined" integral. To clarify, a integral is well defined if either a or b is finite.

$$a = \int_{-\infty}^{0} xf(x)\ dx, b = \int_0^{\infty} xf(x)\ dx.$$

The problem of a "well-defined" integral makes it so that $E(X)$ might exists and be ∞ or $E(X)$ might not exist. An example for when it does not exist is $X \sim$ Cauchy

Expectation rule

$g(x)$ is a random variable if X is a random variable and $g : \mathbb{R} \to \mathbb{R}$ is (measurable) function.

Theorem 5 *If X is a random variable and $g : \mathbb{R} \to \mathbb{R}$ is a (measurable) function, then:*

1. *$Eg(X) = \sum_{x \in X(\Omega)} g(x)p(x)$ if X is discrete with PMF p, $x \in X$*
2. *$Eg(X) \int_{-\infty}^{\infty} g(x)p(x)dx$ if X has density f, when these quantities are well-defined.*

Proof 3 *Let $Y = g(X)$*

1.4 Solomonoff induction

Solomonoff induction is a method for predicting underlying patters in information. When there are multiple choices for a underlying pattern and they fit the data equally, the choice you take should be using Ocam's razor principle. Solomonoff induction is in-fact a formalization around this principle. The principle recommends, when findings explanations to observations, they should be done using the smallest amount of underlying elements (aka less assumptions and the simplest explanation is the best). The name razor is derived from the shaving of unecessary assumptions. The formalization of this induction is something that I might look into later.

2 Information Theory

https://www.inference.org.uk/mackay/itprnn/book.html

Information theory concerns itself with the transmission of information from source to destination. For transmission we want as few bytes of information to pass as possible. Sometimes the transmission may go through a noisy channel that naturally modifies the source. In such cases we are interested in not only transmitting the data but also recovering lost information which will usually add some information used later to recover in case needed. The natural scheme describing such model is:

Source $\to$ Encoder $\to$ (noisy)channel $\to$ Decoder $\to$ Destination

As seen from the arrows we assume one direction, meaning, messages can not be resent. The form above is in-fact not the most common form. Most commonly the encoder and decoder are separated into source encoder, channel encoder and

source decoder, channel decoder. This decoupling of responsibilities as proven by Shannon can be done without loss of efficiency. Because a decoupled system is easier to create and can be done efficiency, when there is a noisy channel it is used. The name of Shannon's revolutionary theorem is the Source-Channel separation theorem.

2.1 Source-Channel separation theorem

def: C is uniquely decodable if C^* is 1-1 (i.e if $x_1 \dots x_n \neq y_1 \dots y_m$ then $C^*(x_1 \dots x_n) \neq C^*(y_1 \dots y_m)$)

2.2 Kraft-McMillan inequality

A B-ary code is a code $C : X \to A^*$ s.t $|A| = B \geq 1$ e.g. $A = \{0, 1\}$ $B = 2$, 2-ary; $A = \{1, 2, 3\}$ $B = 3$, 3-ary

Theorem: (a) McMillan: For any uniquely decodable B-ary code C, $\sum_{x \in X} \frac{1}{B^{l(x)}} \leq 1$ where $l(x) = |C(x)|$ (b) Kraft: If $l : X \to \{1, 2, 3, 4, ...\}$ satisfies $\sum_{x \in X} \frac{1}{B^{l(x)}} \leq 1$ There exists a B-ary prefix code C s.t. $|C(x)| = l(x)$ $\forall x \in X$

Practice:

x	p(x)	C(x)	l(x)
a	0.5	0	1
b	0.25	10	2
c	0.125	110	3
d	0.125	111	3

Proof

2.3 Shannon's entropy equation:

$$H(x) = -\sum_{i=0}^{N-1} p_i \log_2 p_i$$

$\log_2 n = \frac{\log_b n}{\log_b 2}$
numBits $= [H(x)]$
The average minimum number of bites needed to represent a symbol is
$H(x) = -[0.4 \log_2 0.5 + 0.2 \log_2 0.2 + (0.1 \log_2 0.1) * 3]$
$H(x) = -[-0.5 + (-0.46438) + (-0.9965)]$
$H(x) = -[-1.9]$
$H(x) = 1.9$

3 Machine learning

In machine learning we build general tunable systems. When scaled enough, these systems, as showcased by recent advances, have emergent behaviour unexplained by their design. They seem "creative", intelligent, and with character. Properties you can not define on paper concretely. What does it mean to be

creative? How would a intelligent machine look like? How do you define the character of a program or human?

3.1 Defining the problem

It is unlikely that you have a answer ready in your head for these questions. Questions like these and others like them are hard to answer because it is hard to define exactly the concepts they use. They are too abstract and elusive. To combat this problem, machine learning proposes another way of looking at the problem. Instead of trying to define the problem and create a solution based on the problem, why not use computation to understand the problem using data, and solve the problem using this understanding.

This still might be a bit abstract to fully appreciate. And so a more illustrative example might be of better use. Think of two humans. The first one dances and another ties to copy his behaviour. The one copying has all the information of how the first human dances and so he tries to copy his movement. After a couple tries he succeeds and dances exactly like the first human. Now he is asked to copy another dance of the first human. The second time it took him less time since he noticed some regularities in his dances. The third time he uncovered even more patterns. And at some point he was able to copy the first human so well other people have began saying they both have the same dance style.

In this example I try to showcase the presence of regularities and patterns in the dances. Or more generally in data. If all data is unique no information can be compressed and so no style can be formed. It would be so unique it might be called random. For a machine to learn how to form character of speech, or creativity all it has to do it find the patterns that make us think it has character or creativity and make predictions that never go out of the borders of these patterns. These patterns are almost always hidden in plain sight but we are incapable of understanding them because of our limited ability.

This is why, if we can approximate a function to find these patterns in data automatically and generate the appropriate response, we can create a intelligent agent to solve general problems we are not able to define well.

3.2 Learning?

So far as we know the intelligence of a model is linked to it's ability to compress data or, in other words, extract this "hidden" knowledge. Machine Learning shows us a way to extract these features utilizing the vast compute available to us thanks to moors law.

3.3 Deep Learning

Deep learning is a popular technique for creating these general models. Deep learning approximates the original function that "generated" the data through a combination of linear functions. e.g. $f(x) = x^2$ if we sample 10 point of data

from said function, the point of a machine learning model, is to predict this underlying using the points. The initial model or the general function, must contain tunable parameters, through which we can approximate this specific function. $f(\vec{w}, x) = w_0 + w_1 x + w_2 x^2$ where w, is tunable. If we train a model on f we should expect to see w of the form $\vec{w} = (0, 0, 1)$

$$
\begin{aligned}
\text{MSE}(y, \hat{y}) =& \frac{1}{n} \sum_{i=0}^{n} (\hat{y} - y)^2 \\
\hat{y} =& wx + b \\
\text{loss}(y, x) =& (y - wx + b)^2 \\
\frac{\partial l(y, \hat{y})}{\partial w} =& \frac{\partial l(y, \hat{y})}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w} = \frac{1}{n} \sum_{i=0}^{n} 2(\hat{y}_i - y_i) x_i = \frac{2}{n} \sum_{i=0}^{n} x_i (\hat{y}_i - y_i) \\
\frac{\partial l(y, \hat{y})}{\partial b} =& \frac{\partial l(y, \hat{y})}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = \frac{1}{n} \sum_{i=0}^{n} 2(\hat{y}_i - y_i) = \frac{2}{n} \sum_{i=0}^{n} (\hat{y}_i - y_i)
\end{aligned}
$$

4 Excercises

Excercise for Probability Primer measure theory:
Facts Let $(\Omega,\ A,\ P)$ be a probability measure space. $E, F, E_i \in A$

1. $P(E \cup F) = P(E) + P(F)$ if $E \cup F = \emptyset$
2. $P(E \cup F) = P(E) + P(F) - P(E \cap F)$
3. $P(E) = 1 - P(E^c)$
4. $P(E \cap F^c) = P(E) - P(E \cap F)$
5. $P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i) - \sum_{i<j} P(E_i \cap E_j) + \sum_{i<j<k} P(E_i \cap E_j \cap E_k) - \cdots + (-1)^{n+1} P(E_1 \cap E_2 \cap \ldots E_n)$
6. $P(\bigcup_{i=1} E_i) \leq \sum_{n=1} P(E_i)$ and $P(\bigcup_{i=1}^{\infty} E_i) \leq \sum_{n=1}^{\infty} P(E_i)$

Proof 1:

Based on the measure's property of "countable additivity", definition 4 .

Proof 2:

$P(E \cup F) = P(E) + P(F)$ if $E \cup F = \emptyset \implies P(E \cup F) - P(E \cap F)$ if $E \cap F \neq \emptyset$
$\implies P(E \cap F) = P(E) + P(F) + P(E \cup F)$

Proof 3:

$$P(\Omega) = P(E) + P(E^c)$$
$$P(E) = P(\Omega) - P(E^c)$$
$$P(E) = 1 - P(E^c).$$

Proof 4:

if $E \cap F = \emptyset \implies E \cap F^c = E \implies P(E \cap F^c) = P(E)$
if $E \cap F \neq \emptyset \implies E \cap F^c = E \cap F \implies P(E \cap F^c) = P(E \cap F)$
$P(E \cap F^c) = P(E) - P(E \cap F)$.

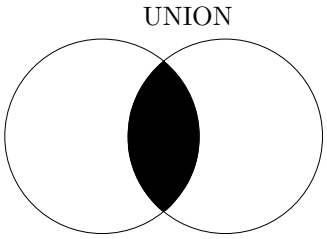


Figure 2: union

Theorem 6 *Chain rule:*

$$\frac{\partial f(g(x))}{\partial x} = \frac{\partial f(g(x))}{\partial g(x)} \frac{\partial g(x)}{\partial x}.$$

Proof 4

$$\frac{\partial f(g(x))}{\partial g(x)} = \frac{f(g(x+h)) - f(g(x))}{g(x+h) - g(x)}, \frac{\partial g(x)}{\partial x} = \frac{g(x+h) - g(x)}{x+h-x}$$

$$\begin{aligned} \frac{\partial f(g(x))}{\partial x} &= \frac{f(g(x+h)) - f(g(x))}{x+h-x} \\ &= \frac{f(g(x+h)) - f(g(x))}{g(x+h) - g(x)} \frac{g(x+h) - g(x)}{x+h-x} \\ &= \frac{\partial f(g(x))}{\partial g(x)} \frac{\partial g(x)}{\partial x}. \end{aligned}$$