1: Information and Probabilistic Modeling

A What is Information?

WRONG ANSWER: Information is Power!

How much information will I reveal when I answer the following questions:

* TH?
* C?T
* ?AT
* What is the next word I am going to say?
* What is the 20th decimal place of π?

You may already have a guesslist of answers, but you are not sure which one is correct. When I answer, I basically reduce that uncertainty:

CO?? → COA?

??	A?
OL	L
AL	T
AT	

Information is reduction of uncertainty

We will justify this principle and build a little theory around it. This theory will help with
- Compression
- communication
- computation
- statistics

→ We seek a quantitative theory of information

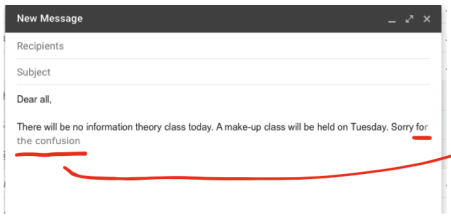
B How do we model uncertainty?

→ Use the number of possible values as a measure of uncertainty

Examples: (1) p is a prime less than 10 ⇒ "uncertainty" = 4

(2) C ? T → ? ∈ {A, O, U} ⇒ "uncertainty" = 2

(3)



This is gmail's auto completion. At this point, perhaps a reasonable list of guesses is

{for the confusion, for the inconvenience}

⇒ uncertainty = 2

(4) x is a 3-dimensional vector within a ball of radius 1.

$\bar{x}$ = x stored using int (16 bits per coordinate)

What is the uncertainty of $\bar{x}$? - Question is not concrete; but try to answer.

(5) p is a prime less than 10^{100} ⇒ "uncertainty" = ?

Each of these example is of different kind.

* Ex 1 is concrete. We may all agree on this
* Ex 2 is not concrete, but based on your vocabulary, you may or may not agree on this
* Ex 3: everyone can perhaps think of other guesses, but they maynot be likely

"for the complication" "for the late notice" "for the pizza"

* Ex4: This one depends on what is an acceptable accuracy for you

* Ex5: This one is very interesting and brings-in computational capabilities

One more point: Suppose we need to store the answer to previous questions. How many bits will it take?

It is the $\log_2$ of our measure of uncertainty.

Some lessons from the discussion above

(a) Uncertainty can vary based on your knowledge/belief

(b) All elements in a guesslist may not be equally likely

(c) There may be a notion of accuracy upto which you may not distinguish your guesses.] not important at this point

→ Use probabilities to model likelihoods of answers

$\mathcal{X}$ ≡ set of possible answers

$|\mathcal{X}| = k$ ≡ no. of possible answers

$p(x)$, $x \in \mathcal{X}$: probability that answer x is correct

* $P = (p(x_1), \ldots, p(x_k))$ is a representation of the guesslist and beliefs

* Who came up with this P? Answerer, general wisdom

* For Ex 4, we need to consider continuous distributions

C Basic Concepts of Probability (Part I: Probability and expectation)

Warning: Little knowledge may be dangerous!

probability mass function, probability density, probability distribution, random variables, expected value, moments, Markov inequality, Chebyshev inequality, Law of Large Numbers, Central Limit Theorem

→ We assume that the audience has some familiarity with these notions. We will only cover the last four.

* Probability mass function (pmf)

For $A \subset \mathcal{X}$, $P(A) = \sum_{x \in A} p(x) = \sum_{x} p(x) \mathbb{1}_A(x)$

($\mathbb{1}_A$: indicator of A)

* Probability densities

e.g. signals with freq. b/w 50 ± 1 Hz

$\mathcal{X} = \mathbb{R}^d$ or other more complicated sets for which you can define:

$$P(A) = \int_A p(x)\,dx = \int p(x)\,\mathbb{1}_A(x)\,dx$$

p: probability density

* Notation

Both definitions of $P(A)$ above have a common expression

$$P(A) = \int \mathbb{1}_A(x)\,dP(x) = \int_A dP(x)$$

$\sum_x \mathbb{1}_A(x)\,p(x)$ for discrete

$\int \mathbb{1}_A(x)\,p(x)\,dx$ for continuous

* Random variables

$\mathbb{1}_A$ we saw above takes values in $\{0,1\}$

- We have defined $\int \mathbb{1}_A(x)\, dP(x) \equiv \int \mathbb{1}_A\, dP$
- Then, we can also define a similar integral for $\sum_{i=1}^{l} a_i \mathbb{1}_{A_i}$ as

$$\int \left(\sum_{i=1}^{l} a_i \mathbb{1}_{A_i}\right) dP \triangleq \sum_{i=1}^{l} a_i \int \mathbb{1}_{A_i}\, dP$$

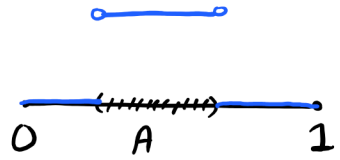
$$= \sum_{i=1}^{l} a_i P(A_i)$$

- We are not allowed to use any subsets A, in general.

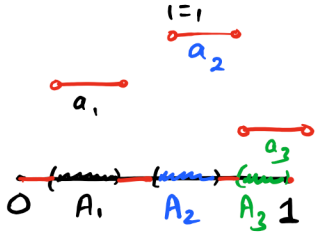
A probability course will teach you which sets are allowed.

- If $\mathcal{X}$ is discrete, all subsets can be allowed
- For $\mathcal{X} = [0,1]$, we can start by allowing all intervals.

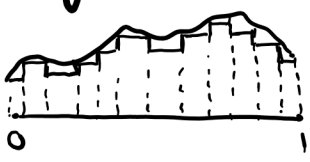
Then, the function $\mathbb{1}_A$ looks like this:



And the function $\sum_{i=1}^{l} a_i \mathbb{1}_{A_i}$ can look like this:



But then "by taking limits" you can have any function:



$$\int f\, dP \triangleq \lim_{n\to\infty} \int f_n\, dP$$

$\downarrow$ of the form $\sum a_i \mathbb{1}_{A_i}$

A random variable is such a function f.

- Takes values in $\mathbb{R}$
- Probabilities are actually on the domain of f
- Denote the r.v by X

The quantity $\int X dP$ is called the expected value of X, and is denoted by $\mathbb{E}[X]$.

- We can talk of probabilities of $X \in B$, $B \subset \mathbb{R}$:

$$P(X \in B) = \mathbb{E}\left[\mathbb{1}_{\{X \in B\}}\right]$$

$$= \int \mathbb{1}_{f^{-1}(B)} \, dP$$

where $f^{-1}(B) = \{x : f(x) \in B\}$.

Examples: Bernoulli, Binomial, Poisson, Geometry, Gaussian Exponential

→ We will model our unknown answers by random variables

Part II: Estimates for random variable

* Markov Inequality

$X \geq 0$ is a nonnegative random variable

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}, \quad \text{for all } a > 0.$$

Proof.

$$\mathbb{E}[X] = \int X \, dP = \int X \left(\mathbb{1}_{\{X>a\}} + \mathbb{1}_{\{X \le a\}} \right) dP$$

$$= \int X \, \mathbb{1}_{\{X>a\}} \, dP + \underbrace{\int X \, \mathbb{1}_{\{X \le a\}} \, dP}_{\ge 0 \text{ since } X \ge 0}$$

$$\ge \int X \, \mathbb{1}_{\{X>a\}} \, dP$$

$$> a \int \mathbb{1}_{\{X>a\}} \, dP$$

$$= a \, \mathbb{P}(X > a). \qquad \blacksquare$$

We can view Markov's inequality as follows:

$$\mathbb{P}\left(X > \frac{1}{\delta} \mathbb{E}[X] \right) \le \delta,$$

that is, an $X \ge 0$ exceeds 10^6 times its expectation with prob. $\le 10^{-6}$.

But how far can X go from $\mathbb{E}[X]$?

$$\mathbb{P}(|X - \mathbb{E}[X]| > t) = \mathbb{P}\left(|X - \mathbb{E}[X]|^2 > t^2 \right)$$

$$\le \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^2 \right]}{t^2} \triangleq \mathrm{Var}(X)$$

That is,

$$\mathbb{P}\left(\mathbb{E}[X] - \sqrt{\frac{\mathrm{Var}(X)}{\delta}} \le X \le \mathbb{E}[X] - \sqrt{\frac{\mathrm{Var}(X)}{\delta}} \right) \ge 1 - \delta.$$

"Chebyshev's inequality"

With prob. $\ge 10^{-6}$, $X \in \mathbb{E}[X] \pm 10^3 \sqrt{\mathrm{Var}(X)}$

Part III: Two useful limit results

* Weak law of large numbers

For independent random variables $X_1, \dots, X_n$,

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i). \qquad \text{(show this)}$$

Therefore, by Chebyshev's inequality

$$\mathbb{P}\left(\left| \sum_{i=1}^n X_i - n\mu \right| > \sqrt{\frac{1}{\delta} \sum_{i=1}^n \mathrm{Var}(X_i)} \right) \le \delta,$$

where $\mu = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i]$.

If $X_1, \dots, X_n$ are iid, we have

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| > \frac{\sigma}{\sqrt{n\delta}} \right) \le \delta.$$

Thus, we can say that $\frac{1}{n} \sum_{i=1}^n X_i$ converges to μ in a sense (this kind of convergence is called "convergence in probability").

Theorem (Weak Law of Large numbers)

For every $\varepsilon > 0$

$$\lim_{n \to \infty} \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| > \varepsilon \right) = 0.$$

(We have a more precise estimate:

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| > \varepsilon \right) \le \delta \quad \text{for every } n \ge \frac{\sigma^2}{\delta \varepsilon^2}$$

* Central limit theorem (Stated without proof)

For iid $X_1, \dots, X_n$ with $\mathbb{E}[X_i] = \mu$, $\mathrm{Var}(X_i) = \sigma^2$,

$$\lim_{n \to \infty} \mathbb{P}\left(\sum_{i=1}^{n} X_i - n\mu > \sqrt{n\sigma^2 t} \right) = Q(t)$$

for every $t > 0$. Here $Q(t)$ denotes the Gaussian-tail probability given by

$$Q(t) := \int_t^\infty \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx \approx e^{-t^2/2}.$$

→ This result gives a more precise estimate of the error distribution in approximating $\sum_{i=1}^{n} X_i$ with $n\mu$. Specifically, the error is roughly $N(0, n\sigma^2)$.

It implies

$$\frac{1}{n} \sum_{i=1}^{n} X_i \approx \mu \pm \sqrt{\frac{2\sigma^2 \log \frac{2}{\delta}}{n}}.$$