Probability for Language Modeling

Part I: Foundations

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Institute of Computer Science
Polish Academy of Sciences

Quantitative Cognitive Linguistics Network
4th July 2024

A (not so) rare specimen of a stochastic parrot

The rise of large language models
with their strengths
(fluent relevant grammatical replies)
and weaknesses
(factual hallucinations)
reopens the old question whether
statistical language modeling
makes sense for language science.

Motivating question

Does statistical language modeling make sense for linguistics?

I think one should replace "Does" with "How can".

There is a related question of a theoretical importance:

How much randomness is there in language and speech ...
... and, precisely, how does it interfere with structure?

This question cannot be answered without a certain understanding of mathematical models of randomness.

These slides provide an intro.

Examples of probability

Particular instances of the common concept:

- Empirical frequency of a repeatable phenomenon. (frequency)
- Limit of such frequencies. (frequency & extrapolation)
- Subjective and evolving belief of a learning agent. (Bayesian)
- Propensity of an unpredictable phenomenon. (randomness)
- Weight in a weighted mean. (abstract)
- Fraction of favorable elementary events. (abstract)
- Relative volume of a figure. (abstract & geometry)
- Result of a smoothing procedure. (computation)
- Output of a complicated black box. (computation)

Having many interpretations & models is good. Let's treat different views as tools and switch from one to another as we learn more!
That's how mathematics works!

What is discrete probability?

Let Ω be a countable set of values (symbols, integers, words, ...).

A probability distribution p is a function of points $\omega \in \Omega$ such that

$$p(\omega) \geq 0, \quad \sum_{\omega \in \Omega} p(\omega) = 1.$$

A probability measure P is a function of events $A \subset \Omega$ such that

$$P(A) = \sum_{\omega \in A} P(\{\omega\}), \quad P(A) \geq 0, \quad P(\Omega) = 1.$$

Obviously, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Simplest example: Uniform measure $P(\{\omega\}) := 1/\#\Omega$.

Example: A cubic die

Define

$$\begin{aligned} P(\{1\}) &= 1/6, & P(\{2\}) &= 1/6, \\ P(\{3\}) &= 1/6, & P(\{4\}) &= 1/6, \\ P(\{5\}) &= 1/6, & P(\{6\}) &= 1/6. \end{aligned}$$

We have

$$\begin{aligned} P(\{\omega : \omega \text{ is odd}\}) &= 1/6 + 1/6 + 1/6 = 1/2, \\ P(\{\omega : \omega \text{ is a square}\}) &= 1/6 + 1/6 = 1/3. \end{aligned}$$

Random variables and expectation

A random variable is a function $X : \Omega \to \mathbb{X}$.

We denote $(X \in B) := \{\omega \in \Omega : X(\omega) \in B\}$.

The expectation of a real random variable $X : \Omega \to \mathbb{X} \in [a, b]$ is

$$\mathbb{E}\, X = \sum_{x \in \mathbb{X}} P(X = x) \cdot x.$$

We can generalize this definition via measure theory
to an uncountable set of values including infinities.

Example: A cubic die

Define

$$P(\{1\}) = 1/6, \qquad P(\{2\}) = 1/6,$$
$$P(\{3\}) = 1/6, \qquad P(\{4\}) = 1/6,$$
$$P(\{5\}) = 1/6, \qquad P(\{6\}) = 1/6.$$

Let $X(\omega) = \omega^2$. We have

$$\begin{aligned} \mathbb{E}\, X &= \sum_{x \in \mathbb{X}} P(X = x) \cdot x \\ &= \sum_{\omega \in \Omega} P(\{\omega\}) \cdot X(\omega) \qquad \text{(for discrete probability only)} \\ &= \sum_{\omega=1}^{6} \frac{\omega^2}{6} = \frac{1 + 4 + 9 + 16 + 25 + 36}{6}. \end{aligned}$$

Indifference + decomposability = independence

Events $A_1, ..., A_n$ are called independent if

$$P(A_1 \cap ... \cap A_n) = P(A_1) \cdot ... \cdot P(A_n).$$

If A is independent of A then either $P(A) = 0$ or $P(A) = 1$.

Random variables $X_1, ..., X_n$ are called independent if

$$P(X_1 \in B_1, ..., X_n \in B_n) = P(X_1 \in B_1) \cdot ... \cdot P(X_n \in B_n).$$

The simplest example is product space $\Omega := \Omega_1 \times ... \times \Omega_n$ and uniform measure $P(\{\omega\}) = 1/\#\Omega$. Then projections $X_1, ..., X_n$, defined as $X_i(\omega) := \omega_i$ for $\omega = (\omega_1, ..., \omega_n)$, are independent.

Example: Five cubic dice

Define

$$P(\{(1,1,1,1,1)\}) = 1/6^5, \qquad P(\{(1,1,1,1,2)\}) = 1/6^5,$$
$$\ldots$$
$$P(\{(6,6,6,6,5)\}) = 1/6^5, \qquad P(\{(6,6,6,6,6)\}) = 1/6^5.$$

Let $X_i(\{(\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}) = \omega_i$. We have

$$\begin{aligned} P(X_1 \text{ is odd}, X_4 \text{ is square}) &= P(X_1 \text{ is odd}) \cdot P(X_4 \text{ is square}) \\ &= \frac{3}{6} \cdot \frac{2}{6} = \frac{1}{6}. \end{aligned}$$

Conditional probability

Conditional probability of A given B is defined as

$$P(A|B) := \frac{P(A \cap B)}{P(B)}.$$

The above definition works only if $P(B) > 0$.

Obviously $P(A|B) = P(A)$ if A and B are independent.

In general, $P(A|B)$ may be smaller or greater than $P(A)$.

But $P(A)$ is a weighted average of conditional probabilities:

$$P(A) = P(B)P(A|B) + P(\Omega \setminus B)P(A|\Omega \setminus B).$$

Example: Five cubic dice

Define

$$P(\{(1,1,1,1,1)\}) = 1/6^5, \qquad P(\{(1,1,1,1,2)\}) = 1/6^5,$$
$$\dots$$
$$P(\{(6,6,6,6,5)\}) = 1/6^5, \qquad P(\{(6,6,6,6,6)\}) = 1/6^5.$$

Let $X_i(\{(\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}) = \omega_i$. We have

$$P(X_1 \text{ is odd}|X_4 \text{ is square}) = P(X_1 \text{ is odd}) = \frac{3}{6} = \frac{1}{2},$$
$$P(X_1 \text{ is odd}|X_1 \text{ is square}) = P(X_1 \text{ is odd and square}) : P(X_1 \text{ is square}) = \frac{1}{6} : \frac{2}{6} = \frac{1}{2}.$$

Accidentally, these two values are the same!

Do theoretical linguists need infinities?

- *Language is an infinite use of finite means.*

  (Wilhelm von Humboldt)
- Potential infinities arise when we are trying to extrapolate finite data, to predict that something happens in an indefinite future, happens ultimately or arbitrarily often.
- Sometimes, working with actual infinities is easier than working with big data. (At least, it can be cheaper.)
- Basically, we need infinities if we work with:
  - sequences of discrete outcomes that can be extended at will,
  - real numbers that can be learned with an arbitrary precision.
- It may be good to be aware of some phenomena that take place in the realm of actual infinities.
- Some of them, connected to ergodic decomposition, have a linguistic interpretation. — **More in further lectures!**

What is measure-theoretic probability?

Measure theory is an extension of discrete probability to an uncountably infinite set of elementary events (infinite sequences of symbols, real numbers, vectors, etc.).

Let Ω be an arbitrary set of values.

A probability measure P is a function of some sets $A \subset \Omega$ s.t.:

- $0 \leq P(A) \leq 1$, $P(\emptyset) = 0$, $P(\Omega) = 1$,
- $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$,
- $P(\bigcup_{n \in \mathbb{N}} A_n) = \sum_{n \in \mathbb{N}} P(A_n)$ if $A_i \cap A_j = \emptyset$. (continuity!)

Simplest example: Lebesgue measure on the unit interval $P(\{r \in \mathbb{R} : a \leq r \leq b\}) := b - a$ for $0 \leq a \leq b \leq 1$.

CAVEAT: In this case, we have

$$P(\{\omega\}) = 0, \quad P(A) \neq \sum_{\omega \in A} P(\{\omega\}).$$

Moreover, P may be not determined for some sets $A \subset \Omega$.

Learning measure theory

- The framework of measure theory is quite heavy.
- It takes time to absorb it. (→ personal accounts)
- You can learn it on your own but a good teacher speeds it up.
- Patience and goal-oriented motivation matter, too.
- Dunning-Kruger effect:
  The first impression usually is that you lose confidence in apparently elementary reasonings that turn out not to be so (high school integrals, limits, foundations of math, etc.).
- Once you build correct intuitions, you realize that a problem to solve usually relies on a few important theorems.
- Some opaque definitions are opaque because they are such so as to circumvent a few annoying counterexamples.

What is it good for? An example

Conditional probability of A given B is defined as

$$P(A|B) := \frac{P(A \cap B)}{P(B)}.$$

The above definition works only if $P(B) > 0$.

In measure-theoretic probability, we also consider limits

$$P(X_0|X_1, X_2, \ldots) = \lim_{n\to\infty} P(X_0|X_1, \ldots, X_n) \tag{1}$$

for $\lim_{n\to\infty} P(X_1, \ldots, X_n) = 0$. The standard way of introducing so generalized conditional probabilities is through Radon-Nikodym derivatives (densities of measures). Then convergence (1) follows by the martingale convergence theorem.

Another example: Fair-coin or Bernoulli process

A sequence of independent coin flips (bits, spins) with a uniform distribution is the simplest model of a random phenomenon.

Formally, we have $Z_1^n := (Z_1, Z_2, ..., Z_n)$ such that

$$P(Z_1^n = z_1^n) = 2^{-n}, \quad z_1^n \in \{-1, 1\}^n .$$

We have $\mathbb{E}\, Z_i = 0$ and $\operatorname{Var} Z_i := \mathbb{E}\, Z_i^2 - (\mathbb{E}\, Z_i)^2 = 1$.

This simple stochastic process exhibits a few important laws:

1. Law of large numbers (LLN):
   The sample mean approaches its expectation,
   $\lim_{n\to\infty} \sum_{i=1}^n Z_i/n = 0$ with probability 1.
2. Central limit theorem (CLT):
   The distribution of rescaled sample mean $\sum_{i=1}^n Z_i/\sqrt{n}$ approaches the Gauss distribution $N(0,1)$ as $n \to \infty$.
3. Law of the iterated logarithm (LIL):
   $\limsup_{n\to\infty} \left|\sum_{i=1}^n Z_i\right| / \sqrt{2n \ln \ln n} = 1$ with probability 1.

1. Probability
2. Measure
3. Computation
4. Information

The need of a stronger concept

Assume independent coin flips with a uniform distribution.

Consider two finite binary sequences:

01011010110110100101010100010

11111111111111111111111111111

Both have the same probability but only the first one may have the statistical properties of a fair coin. Why???

We crave for a concept of randomness that would apply to particular sequences rather than their ensembles.

Ideally, a random sequence would be a particular sequence that exhibits all statistical laws of the uniform independent coin flips.

Such sequences exist in the Platonic world.
They can be defined via the theory of computation.

Register machine

An approach equivalent to Turing machines.

A computer is the interpreter of this programming language:
– Variables $R_1, R_2, R_3, \ldots$ take values in natural numbers.
– A program is a finite list of commands of form:

- $R_j := 0;$
- $R_j := R_j + 1;$
- $R_j := R_k;$
- IF $R_j = R_k$ GOTO m;

A computable function is a function computed by some program.

Fixing a 1-to-1 mapping between numbers and strings, we can speak of computable functions for string arguments and values.

Halting problem

Computable functions can be enumerated by listing all programs:

$$F_1, F_2, F_3, \dots$$

The computation of F_n on argument m either loops or halts.

The universal function $U(n, m) := F_n(m)$ is computable.

An example of a not computable function is the halting function

$$H(n, m) := \begin{cases} \text{“halts”} & \text{if } F_n(m) \text{ halts,} \\ \text{“loops”} & \text{if } F_n(m) \text{ loops.} \end{cases}$$

If it were computable, we might define a computable function

$$F_k(m) \text{ that } \begin{cases} \text{halts} & \text{if } H(m, m) = \text{“loops”}, \\ \text{loops} & \text{if } H(m, m) = \text{“halts”}. \end{cases}$$

and obtain contradiction “loops” $= H(k, k) =$ “halts”.

Kolmogorov complexity

Let x be a discrete object (string, natural number etc.).

The Kolmogorov complexity $K(x)$ of object x is
the minimal length of a binary program that computes x.

Function $x \mapsto K(x)$ is not computable (by the halting problem)
but it is semi-computable (if $K(x) \leq m$ then we can prove the validity of this relation in a finite time).

Algorithmic randomness

- A particular sequence of coin flips is called (algorithmically) random when the shortest program that prints out this sequence is almost as long as the sequence.
(Random sequences cannot be compressed.)
- The fraction of random sequences among all binary sequences tends to 1 as we take longer and longer sequences.
(Almost all binary sequences are random.)
- One cannot prove that a given sequence is random but one can prove that a sequence isn't random if it isn't random.
(We cannot provably generate a random sequence.)
- If a given computable property is satisfied for almost all sequences then it holds for exactly all random sequences.
(Probability theory describes properties of random sequences.)

In physical data, probably only non-random sequences exist!
But proving non-randomness of some may take a formidable time.

Resource-bounded randomness

Resource-bounded randomness theory considers weaker notions:

A martingale is a function $d : \{-1, 1\}^* \to [0, \infty)$ s.t.

$$d(z_1^n) = \frac{Q(Z_1^n = z_1^n)}{P(Z_1^n = z_1^n)} = 2^n Q(Z_1^n = z_1^n),$$

where Q is some probability measure.

We say that:

- d fails on $(z_n)_{n\in\mathbb{N}}$ if $\sup_{n\in\mathbb{N}} d(z_1^n) < \infty$.
- d is $f(n)$-time if $d(z_1^n)$ can be computed in time $f(n)$.
- $(z_n)_{n\in\mathbb{N}}$ is $f(n)$-random if any $f(n)$-time d fails on $(z_n)_{n\in\mathbb{N}}$.

Schnorr's theorem

If $(z_n)_{n\in\mathbb{N}}$ is n^2-random then $(z_n)_{n\in\mathbb{N}}$ satisfies LLN.

If a fixed sequence cannot be essentially quickly compressed then it satisfies some particular laws of randomness!

The need of a more effective approach

Informally speaking, we have the equivalence

$$\text{information} = \text{novelty} = \text{unpredictability} = \text{randomness}.$$

Kolmogorov complexity, i.e., the length of the shortest program to generate a string, is an uncomputable measure of information.

We need a concept of information that would be approximately equal to Kolmogorov complexity but could be effectively computed.

Such a concept can be defined and is called Shannon entropy.

Shannon entropy

We consider a random variable $P(X) : \Omega \to [0,1]$ such that

$$P(X)(\omega) := P(X = x) \text{ if } X(\omega) = x.$$

The Shannon entropy of a random variable $X : \Omega \to \mathbb{X}$ is

$$\begin{aligned} H(X) &:= \mathbb{E}\left[-\log_2 P(X)\right] \\ &= -\sum_{x \in \mathbb{X}} P(X = x) \log_2 P(X = x) \in [0, \log \# \mathbb{X}]. \end{aligned}$$

It is a measure of uniformity of a distribution:

$$\begin{aligned} H(X) = \log \# \mathbb{X} &\iff P(X = x) = 1/\# \mathbb{X}, \\ H(X) = 0 &\iff P(X = x) \in \{0, 1\}. \end{aligned}$$

The Shannon entropy can be infinite if $\mathbb{X}$ is infinite.

Conditional entropy

We consider a random variable $P(X) : \Omega \to [0, 1]$ such that

$$P(X|Y)(\omega) := P(X = x|Y = y) \text{ if } (X, Y)(\omega) = (x, y).$$

The conditional entropy of $X : \Omega \to \mathbb{X}$ given $Y : \Omega \to \mathbb{Y}$ is

$$\begin{aligned} H(X) &:= \mathbb{E}\left[-\log_2 P(X|Y)\right] = H(X, Y) - H(Y) \\ &= \mathbb{E}\left[-\sum_{x \in \mathbb{X}} P(X = x|Y) \log_2 P(X = x|Y)\right] \in [0, H(X)]. \end{aligned}$$

It is a measure of unpredictability of a random variable:

$$\begin{aligned} H(X|Y) = H(X) &\iff P(X, Y) = P(X)P(Y), \\ H(X|Y) = 0 &\iff X = f(Y). \end{aligned}$$

Mutual information

The mutual information between $X : \Omega \to \mathbb{X}$ and $Y : \Omega \to \mathbb{Y}$:

$$I(X;Y) := \mathbb{E}\left[\log_2 \frac{P(X,Y)}{P(X)P(Y)}\right] = H(X) + H(Y) - H(X,Y)$$
$$= H(X) - H(X|Y) \in [0, \min\{H(X), H(Y)\}].$$

It is a measure of dependence of distributions:

$$I(X;Y) = H(X) \iff X = f(Y),$$
$$I(X;Y) = H(Y) \iff Y = g(X),$$
$$I(X;Y) = 0 \iff P(X,Y) = P(X)P(Y).$$

Prefix-free codes

A prefix-free code is a mapping $B : \mathbb{X} \to \{0,1\}^*$ such that

$$B(x) = B(y)u \implies x = y \text{ and } u = \lambda.$$

Then we can decipher the objects from codeword concatenation:

$$B(x_1)...B(x_n) = B(y_1)...B(y_m) \implies (x_1, ..., x_n) = (y_1, ..., y_m).$$

An incomplete distribution is a mapping $q : \mathbb{X} \to [0,1]$ such that

$$\sum_{x \in \mathbb{X}} q(x) \leq 1.$$

By the Kraft inequality, for each prefix-free code B, function $q(x) = 2^{-|B(x)|}$ is an incomplete distribution. Conversely, for each incomplete distribution Q there is a prefix-free code B, called the Shannon-Fano code, such that $|B(x)| = \lceil -\log q(x) \rceil$.

Incomplete distributions

Incomplete distributions satisfy two important inequalities:

- non-negativity of relative entropy:

$$\sum_{x \in \mathbb{X}} p(x) \log \frac{p(x)}{q(x)} \geq 0.$$

- Barron's inequality:

$$\sum_{x \in \mathbb{X}} p(x) 1 \left\{ \log \frac{p(x)}{q(x)} \leq -m \right\} \leq 2^{-m}.$$

As a result, each prefix-free code satisfies

$$\mathbb{E} |B(X)| = \sum_{x \in \mathbb{X}} P(X = x) |B(X = x)| \geq H(X),$$

$$P(|B(X)| \leq -\log P(X) - m) \leq 2^{-m}.$$

In particular, the Shannon-Fano code for $P(X = \cdot)$ satisfies

$$H(X) \leq \mathbb{E} |B(X)| \leq H(X) + 1.$$

Kolmogorov complexity and Shannon entropy

The following holds for some version of Kolmogorov complexity.

The minimal program that computes x is a prefix-free code for x.

Hence the Kolmogorov complexity, its length, satisfies

$$\mathbb{E}\, K(X) := \sum_{x \in \mathbb{X}} P(X = x) K(x) \geq H(X),$$

$$P(K(X) \leq -\log P(X) - m) \leq 2^{-m}.$$

If the Shannon-Fano code for $P(X = \cdot)$ can be decoded by a program of length C then Kolmogorov complexity is bounded by

$$K(x) \leq C - \log P(X = x).$$

Hence $H(X) \leq \mathbb{E}\, K(X) \leq H(X) + C$.

For random distributions of X, constant C can be arbitrarily large.

What is this all good for?

1. **Discrete probability:**
   One cannot simply state large language models without that!
2. **Measure theory:**
   Is indispensable to discuss asymptotic properties of statistical language models. Although it's a complicated tool, its applications inform a theoretical linguistic point of view.
3. **Kolmogorov complexity:**
   Is necessary to make the concept of randomness agree with our preformal intuitions. Although it is uncomputable, it allows to speak of information content of a single text.
4. **Shannon entropy:**
   Is a measure of information that may be sometimes easier to estimate than the Kolmogorov complexity. It is a smoothed out version of Kolmogorov complexity that applies to ensembles of texts.

Further reading

P. Billingsley. *Probability and Measure*. Wiley & Sons, 1979.

T. M. Cover and J. A. Thomas. *Elements of Information Theory, 2nd ed.* Wiley & Sons, 2006.

Ł. Dębowski. A short course in universal coding. Book manuscript in preparation, 2024.

M. Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications, 3rd ed.* Springer, 2008.

J. von Plato. *Creating modern probability: Its mathematics, physics, and philosophy in historical perspective*. Cambridge University Press, 1994.