Lecture notes from
Models and Numerical Methods

https://github.com/Grufoony/Physics_Unibo

Contents

1 Resume of measure theory

We need to define a mathematical model that generates sequences from an *alphabet* $\mathcal{A}$, which can be any finite set. We will denote both set of finite and infinite sequences as $\mathcal{A}^* = \cup_{n\in\mathbb{N}}\mathcal{A}^n$ and $\mathcal{A}^{\mathbb{N}}$. Now we can define a sequence, or a *word*, $\omega \in \mathcal{A}^n$ and denote with $|\omega| = n$ its length. In particular, we will use the notation $\omega_i^j = (\omega_i, \dots, \omega_j)$. We can also take $\mathcal{A}^{\mathbb{Z}}$ as two-sided alphabet.

Definition 1. *A **measurable space** $(\Omega, \mathcal{F})$ is (usually) defined by a compact metric space Ω and a σ-algebra $\mathcal{F}$.*

We will denote the canonical cylinder on Ω as $[a_1^n] = \{y \in \Omega \mid y_1 = x_1, \dots, y_n = x_n\}$. To figure out that this is actually a cylinder, let's pretend to take (r, ϕ, h) cylindrical coordinates, fixing the radius $r = r_0$, letting the angle and the height free. In other words, a cylinder contains all sequences which start with a given word.

Our space has a topology, so we can take $\mu \approx \mathrm{m}$ (metric) absolutely continuous w.r.t. the Lebesgue measure on $\Omega = \mathbb{R}^n$. So it exists $\varphi \in \mathbb{L}^1(m)$ such that $\mu(f) = \int dm f(m)\varphi(m)$. Consider now the function

$$g_{\mathcal{A}}(z, z') = \begin{cases} 1 & z = z' \\ 0 & z \neq z' \end{cases} \qquad \forall z, z' \in \mathcal{A}$$

Taking $x, y \in \Omega$ infinite sequences it is possible to prove that

$$\widetilde{d}(x, y) = \sum_{n=1}^{\infty} 2^{-n} g_{\mathcal{A}}(x_n, y_n)$$

is a metric over Ω. Taking $x^{(n)} \in \Omega$ sequence of infinite sequences, given $0 < \lambda = \frac{1}{|A|} < 1$, we have that $d(x, y) = \lambda^{n(x,y)} \quad \forall x, y \in \Omega$ is also a metric over Ω, with $n(x, y) = \min\{k | x_k \neq y_k\}$. Moreover, d and $\widetilde{d}$ define the same topology. The open balls are, $\forall x \in \Omega, \quad r > 0$

$$\mathcal{B}(x, r) = \{y \in \Omega \mid d(x, y) \leq r\} = \left\{ y \in \Omega \mid x_k = y_k \ \forall\ 1 \leq k \leq \frac{\ln r}{\ln \lambda} \right\}$$

Definition 2. *$\mathcal{F}$ is a **Borel σ-algebra** if is a set of subsets of Ω such that $\Omega \in \mathcal{F}$*

So a σ-algebra is actually a collection of all measurable sets.

Definition 3. *The **space of probability measures** is $\mathcal{P}(\Omega) = \{\mu \mid \mu(\Omega) = 1\}$*

Taking now a map $T : \Omega \to \Omega$ with a probability measure μ, μ is called **T-invariant** if $\forall A \in \mathcal{F}, \mu(T^{-1}(A)) = \mu(A)$. In other words, the push-forward measure as to satisfy $T * \mu \equiv \mu \cdot T^{-1} = \mu$. We will denote the space of invariant probability measures with $\mathcal{P}_{\mathcal{F}} \subset \mathcal{P}$.

The **shift** map $\sigma : \Omega \to \Omega$ such that $\sigma(x_0, x_1, \ldots, x_n, \ldots) = \sigma(x_1, \ldots, x_n, \ldots)$ represents an important map over our space. Notice that the shift map could represent our time, helping us to code a dynamical system (more details in the next sections). Moreover, one can verify that if μ is i.i.d. then it's shift-invariant, i.e. $\mu(\sigma^{-1}([x_1^n])) = \mu([x_1^n])$ and $\mu\left(\left[x_{n+k}^{m+k}\right]\right) = \mu([x_n^m])$.

Theorem 1 (Borel-Cantelli lemma)**.**

$$\{E_n\} \mid \sum_{n=1}^{\infty} \mu(E_n) < \infty \implies \mu(\limsup E_n) = 0$$

This lemma states that, beside null-measure sets, typically an $x \in \Omega$ only belongs to finitely many E_k's.
In order to clarify, let's take $E_m \subset \Omega$ such that $x \in \limsup E_n \Leftrightarrow x \in \cup_{m=n}^{\infty} E_m \ \forall n$, then it exists $n_j \to \infty$ sequence such that $x \in E_{n_j} \ \forall j$. Consequently, E_n are becoming rare as n increases.

Definition 4. *Let's consider a sequence of events* $\{E_1, \ldots, E_n\}$*. These are* ***independent*** *if* $\mu(E_1 \cap \ldots \cap E_n) = \mu(E_1) \ldots \mu(E_n)$

2 Stochastic Processes

Definition 5. *A **stochastic process** is an infinite sequence of random variables X_n with values in $\mathcal{A}$ defined by the k^{th} order joint distribution:*

$$\mu_k\left(a_1^k\right) = \mathbb{P}\left(X_1^k = a_1^k\right) \quad a_1^k \in \mathcal{A}$$

We need also a consistency condition:

$$\mu_t\left(a_1^t\right) = \sum_{a_0 \in \mathcal{A}} \mu_{t+1}\left(a_0^t\right) = \sum_{a_{t+1} \in \mathcal{A}} \mu_{t+1}\left(a_1^{t+1}\right)$$

Equivalently, we can define a stochastic process through the conditional probability

$$\mu\left(a_t | a_1^{t-1}\right) = \frac{\mu_t\left(a_1^t\right)}{\mu_{t-1}\left(a_1^{t-1}\right)}$$

The μ_k are called **marginals** and, in order to be a probability, they must satisfy the normalization condition

$$\sum_{a_1^k \in \mathcal{A}} \mu_k\left(a_1^k\right) = 1$$

We notice that this sum is exponentially growing in k, so it's impossible to approximate the measure.

Definition 6. *A stochastic process is **stationary** if*

$$\mu\left(a_1^k\right) = \mu\left(a_{t+1}^{t+k}\right) \quad \forall a_1^\infty \in \mathcal{A}^{\mathbb{N}}$$

Definition 7. *An **information source** is a stationary, ergodic, stochastic process.*

Definition 8. *A process or a source is a **shift-invariant Borel probability measure** μ on the topological space $\mathcal{A}^{\mathbb{Z}}$ of doubly-infinite sequences $x = \{x_n\}_{n \in \mathbb{Z}}$, drawn from a finite (i.e. countable) alphabet $\mathcal{A}$*

Furthermore, it is trivial that we can write any standard cylinder as

$$\left[x_1^t\right] = \sqcup_{a \in \mathcal{A}} \left[x_1, \ldots, x_t, a\right]$$

It's easy to check that

$$\mu \in \mathcal{P}_I(\Omega) \mid \mu \circ \sigma^{-1} = \mu \Leftrightarrow \sum_{a \in \mathcal{A}} \mu_{t+1}\left(a, x_1, \ldots, x_t\right) = \mu_t\left(x_1^t\right)$$

Neural networks are heuristically approximating μ.

Theorem 2 (Kolmogorov representation theorem)**.** *If $\{\mu_n\}$ is a sequence of measure defining a process then there is a unique Borel probability measure μ on $\mathcal{A}^\infty$ such that, $\forall k \geq 1$ and $\forall \left[a_1^k\right]$ cylinder*

$$\mu\left(\left[a_1^k\right]\right) = \mu_k\left(a_1^k\right)$$

2.1 Markov's Models

Markov's model is a stochastic model used to model pseudo-randomly changing systems. In a Markov's process the n element probability depends only on previous k-elements

$$\mu\left(x_n \mid x_0, x_1, \ldots, x_{n-1}\right) := \mu\left(x_n \mid x_k, x_{k+1}, \ldots, x_{n-1}\right)$$

A **Markov's chain** is a Markov's process where the n element depends only on the current state ($n-1$ element). For this reason a Markov's chain is a memoryless process (present time process).

$$\mu\left(x_n \mid x_0, x_1, \ldots, x_{n-1}\right) := \mu\left(x_n \mid x_{n-1}\right)$$

We can define a **Markov's measure**. Let's call $\mathbf{p} = (p_1,\ p_2, \ldots,\ p_l)$ the probability vector that gives us the probability that a character is extracted from the alphabet $\mathcal{A}$ and $P = [p_{ij}]$ the $l \times l$ matrix that describe the probability than the j character is extracted when the previous one is the i. We know that $\mathbf{p}$ is normalized and P is a stochastic matrix

$$p_j \geq 0 \quad \sum_{j=1}^{l} p_j = 1 \qquad \sum_{h=1}^{j} p_{ij} = 1\ \forall i$$

Since P is stochastic it has the unit vector as eigenvector with 1 as eigenvalue. So for the *Pearson-Frobenius theorem* all the eigenvalues are contained inside the complex circle with radius 1. We say that $\mathbf{p}$ is *invariant* if it's a P's eigenvector. We can define for all n the *Markov's measure*

$$\mu_n\left(x_1,\ \ldots\ ,\ x_n\right) = p_{x_1} P_{x_1 x_2} P_{x_2 x_3} \ldots P_{x_{n-1} x_n}$$

2.2 Hidden Markov's Models

Having an HMM (hidden Markov's model) implies having some hidden states, which we know that exist and some observable which we can actually measure. Depending on the data we got we can have three different types of problems:

- *likelihood*, i.e. determine the likelihood function $\mathcal{L}\left(\mathbf{x}|\theta\right)$
- *decoding*, i.e. discover the best hidden sequence Q given $\mathbf{x}$ and θ
- *learning*, i.e. given $\mathbf{x}$ and the hidden states Q try to infer θ

Let's consider $\mathcal{A} = \{o_1, \ldots, o_n\}$ our set of observable and $\mathcal{B} = \{q_1, \ldots, q_l\}$ our hidden states, with $\mathbf{p} = (p_x)_{x \in \mathcal{B}}$ and $P = [P_{x,y}]_{x,y \in \mathcal{B} \times \mathcal{B}}$. Let's assume it exists the stochastic matrix (on the rows) $R = [R_{y,x}]_{y \in \mathcal{B}, x \in \mathcal{A}}$ such that $\forall y \in \mathcal{B}\ \sum_{x \in \mathcal{A}} R_{y,x} = 1$. Then, our measure is

$$\mu_t\left(x1, \ldots, x_t\right) = \sum_{(y_1, \ldots, y_t) \in \mathcal{B}} \mathbf{p}_{y_1} R_{y_1, x_1} P_{y_1, y_2} \ldots P_{y_t - 1, y_t} R_{y_t, x_t}$$

3 Entropy

The first definition of entropy (of a random variable) was given by Shannon:

Definition 9. *Let X be a random variable which takes values in $\mathcal{A} = \{a_1, \ldots, a_k\}$ with probabilities $\mu_j = \mathbb{P}(X = a_j)$ then its **entropy** its*

$$H(X) = \sum_{j=1}^{k} \mu_j \log \mu_j$$

Definition 10. *The **n-block entropy** is defined as*

$$H_n(\mu) = -\sum_{|\omega|=n} \mu(\omega) \log \mu(\omega)$$

We can see the n-block entropy as the entropy of a word of length n, i.e. $H(X_1^n)$. The first question we can arise is: how much information do I gain adding one character to the sequence?

Definition 11. *The **entropy rate** is defined as $h_n(\mu) = H_{n+1}(\mu) - H_n(\mu)$*