UNIVERSIDAD DE GRANADA

MUTUAL INFORMATION IN UNSUPERVISED MACHINE LEARNING

FRANCISCO JAVIER SÁEZ MALDONADO

Bachelor's Thesis

Computer Science and Mathematics

Tutor

Nicolás Pérez de la Blanca Capilla

FACULTY OF SCIENCE

H.T.S. OF COMPUTER ENGINEER AND TELECOMMUNICATIONS

Granada, Thursday 25th February, 2021

ABSTRACT

Abstract

CONTENTS

Part I

BASIC CONCEPTS

In this part we will introduce the underlying concepts of probability theory and probability distributions that will be needed.

1

PROBABILITY

Underneath each experiment involving any grade of uncertainty there is a *random variable*. This is no more than a *measurable* function between two *measurable spaces*. A probability space is composed by three elements: $(\Omega, \mathscr{A}, \mathcal{P})$. We will define those concepts one by one.

1.1 BASIC NOTIONS

Definition 1. Let Ω be a non empty sample space. $\mathscr{A}$ is a $\sigma-$algebra over Ω if it is a family of subsets of Ω that verify that the emptyset is in $\mathscr{A}$, and it is closed under complementation and countable unions. That is:

- $\emptyset \in \mathscr{A}$.
- If $A \in \mathscr{A}$, then $\Omega \backslash A \in \mathscr{A}$.
- If $\{A_i\}_{i\in\mathbb{N}} \in A$ is a numerable family of $\mathscr{A}$ subsets, then $\cup_{i\in\mathbb{N}} A_i \in \mathscr{A}$.

The pair $(\Omega, \mathscr{A})$ is called a *measurable space* To get to our probability space, we need to define a *measure* on the *measurable space*.

Definition 2. Given $(\Omega, \mathscr{A})$ a measurable space, a *measure* $\mathcal{P}$ is a countable additive, non-negative set function on this space. That is: $\mathcal{P} : \mathscr{A} \to \mathbb{R}_0^+$ satisfying:

- $\mathcal{P}(A) \geq \mathcal{P}(\emptyset) = 0$ for all $A \in \mathscr{A}$,
- $P(\cup_n A_n) = \sum_n P(A_n)$ for any countable collection of disjoint sets $A_n \in \mathscr{A}$.

If $\mathcal{P}(\Omega) = 1$, $\mathcal{P}$ is a *probability measure* or simply a *probability*. With the concepts that have just been explained, we get to the following definition:

Definition 3. A *measure space* is the tuple $(\Omega, \mathscr{A}, \mathcal{P})$ where $\mathcal{P}$ is a *measure* on $(\Omega, \mathscr{A})$. If $\mathcal{P}$ is a *probability measure* $(\Omega, \mathscr{A}, \mathcal{P})$ will be called a *probability space*.

Throughout this work, we will be always in the case where $\mathcal{P}$ is a probability measure, so we will always be talking about probability spaces. Some notation for these measures must be introduced. Let A and B be two events. The notation $P(A, B)$ reffers to the probability of the intersection of the events A and B, that is: $P(A, B) := P(A \cap B)$. It is clear that since $A \cap B = B \cap A$, then $P(A, B) = P(B, A)$. We remark the next definition since it will be important.

Definition 4. Let A, B be two events in Ω. The *conditional probability* of B given A is defined as:

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

There is an alternative way to state the definition that we have just made.

Theorem 1 (Bayes' Theorem)**.** *Let A, B be two events in Ω, given that $P(B) \neq 0$. Then*

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

Proof. Straight from the definition of the conditional probability we obtain that:

$$P(A, B) = P(A|B)P(B).$$

We also see from the definition that

$$P(B, A) = P(B|A)P(A)$$

Hence, since $P(A, B) = P(B, A)$,

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

□

However, events might not give any information about another event occurring. When this happens, we call those events to be *independent*. Mathematically, if A,B are independent events:

$$P(A, B) = P(A)P(B)$$

and as a consequence of this, the conditional probabilty of those events is $P(A|B) = P(A)$. For a finite set of events $\{A_i\}_{i=1}^{n}$, we say that they are mutually independent if and only if every event is independet of any intersection of the other events. That is, if $\{B_i\} \subset \{A_i\}$, then

$$P\left(\cap_{i=1}^{k} B_i\right) = \prod_{i=1}^{k} P(B_i) \qquad \text{for all } k \leq n$$

Random variables (R.V.) can now be introduced. Their first property is that they are measurable functions. This kind of functions are defined as it follows:

Definition 5. Let $(\Omega_1, \mathscr{A}), (\Omega_2, \mathcal{B})$ be measurable spaces. A function $f : \Omega_1 \to \Omega_2$ is said to be *measurable* if, $f^{-1}(B) \in \mathscr{A}$ for every $B \in \mathcal{B}$.

As a quick note, we can affirm that if f, g are real-valued measurable functions, and $k \in \mathbb{R}$, it is true that kf, $f + g$, fg and f/g (if g is not the identically zero function) are also *measurable functions*.

We are now ready to define one of the concepts that will lead us to the main objective of this thesis.

Definition 6 (Random variable)**.** Let $(\Omega, \mathscr{A}, \mathcal{P})$ be a probability space, and $(E, \mathcal{B})$ be a measurable space. A *random variable* is a measurable function $X : \Omega \to E$, from the probability space to the measurable space. This means: for every subset $B \in (E, \mathcal{B})$, its preimage

$$X^{-1}(B) = \{\omega : X(\omega) \in B\} \in \mathscr{A}.$$

Using that sums, products and quotients of measurable functions are measurable functions, we obtain that *sums, products and quotients of random variables are random variables*.

Let now X be a R.V. The *probability* of X taking a concrete value on a measurable set contained in E, say, $S \in E$, is written as:

$$P_X(S) = P(X \in S) = P(\{a \in \Omega : X(a) \in S\})$$

A very simple example of random variable is the following:

Example 1. Consider tossing a coin. The possible outcomes of this experiment are *Heads or Tails*. Those are our random events. We can give our random events a possible value. For instance, let *Heads* be 1 and *Tails* be 0. Then, our random variable looks like this:

$$X = \begin{cases} 1 & \text{if we obtain heads} \\ 0 & \text{if we obtain tails} \end{cases}$$

In the last example, our random variable is *discrete*, since the set $\{X(\omega) : \omega \in \Omega\}$ is finite. A *Random Variable* can also be *continuous*, if it can take any value within an interval.

1.2 EXPECTATION OF A RANDOM VARIABLE

Definition 7. The *cumulative distribution function* F_X of a real-valued random variable X is its probability of taking value below or equal to x. That is:

$$F_X(x) = P(X \leq x) = P(\{\omega : X(\omega) \leq x\}) = P_X((-\infty, x]) \quad \forall x \in \mathbb{R}$$

We can difference between certain types of random variables. If the image , $\mathcal{X}$, of X is countable, we call it a *discrete* random variable. Its *probability mass function* p gives the probability of the R.V. being equal to a certain value:

$$p(x) = P(X = x).$$

If the comulative distribution function of our random variable X is continuous everywhere, then X is a *continuous* random variable. In this case there might exist a non-negative Lebesgue-integrable function f such that:

$$F_X(x) = \int_{\infty}^{x} f(t)dt,$$

called the *probability density function* of X.

We are now ready to introduce the *expectation* of a random variable. Imagine observing a wide number of outcomes from our random variable, and take the average of these random values. The expectation is the value of this average when we take *infinite* outcomes of our random variable.

Definition 8 (Expectation of a *R.V.*)**.** Let X be a non negative random variable on a probability space $(\Omega, \mathscr{A}, \mathcal{P})$. The expectation $E[X]$ of X is defined as:

$$E[X] = \int_{\Omega} X(\omega)\, dP(\omega).$$

The expectation of a random variable will be also denoted as μ. Now, if X is generic $R.V$, the expectation is defined as:

$$E[X] = E[X^+] - E[X^-]$$

where X^+, X^- are defined as it follows:

$$X^+(\omega) = \max(X(\omega), 0), \qquad X^-(\omega) = \min(X(\omega), 0).$$

The expectation $E[X]$ of a random variable is a linear operation. That is, if Y is another random variable, and $\alpha, \beta \in \mathbb{R}$, then

$$E[\alpha X + \beta \mathcal{Y}] = \alpha E[X] + \beta E[\mathcal{Y}].$$

This is a trivial consequence of the linearity of the *Lebesgue integral*.

As a note, if X is a *discrete* random variable and $\mathcal{X}$ is its image, its expectation can be computed as:

$$E[X] = \sum_{x \in \mathcal{X}} x P_X(x),$$

where x is each possible outcome of the experiment, and $P_X(x)$ the probability under the distribution of X of the outcome x. The expression given in Def. **??** 8 before generalizes this particular case.

Using the definition of the *expectation* of a random variable, we can approach to the *moments* of a random variable.

Definition 9. If $k \in \mathbb{N}$, then $E[X^k]$ is called the $k - th$ moment of X.

If we take $k = 1$, we have the definition of the *expectation*. It is sometimes written as $m_X = E[X]$, and called the *mean*. We use the *mean* in the definition of the variance:

Definition 10. Let X be a random variable. If $E[X^2] < \infty$, then the *variance* of X is defined to be

$$\operatorname{Var}(X) = E[(X - m_X)^2] = E[X^2] - m_X^2.$$

Thanks to the linearity of the *expectation* of a random variable, it is easy to see that, if $a, b \in \mathbb{R}$, then

$$\operatorname{Var}(aX + b) = E[(aX + b) - E[aX + b])^2] = a^2 E[(X - m_X)^2] = a^2 \operatorname{Var}(X).$$

Usually, when it comes to applying these concepts to a real problem, we will be looking at multiple variables. We would like to have a collection of random variables each one representing one of this variables. In order to set the notation for these kinds of situations, we will introduce *random vectors*.

Definition 11. A random vector is a row vector $\mathbf{X} = (X_1, \dots, X_n)$ whose components are real-valued random variables on the same probability space $(\Omega, \mathscr{A}, P)$.

The probability distribution of a random variable can be extended in to the *joint probability distribution* of a random vector.

Definition 12. Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random vector. The *cumulative distribution funcion* $F_{\mathbf{X}} : \mathbb{R}^n \to [0, 1]$ of $\mathbf{X}$ is defined as:

$$F_{\mathbf{X}}(x) = P(X_1 \leq x_1, \dots, X_n \leq x_n).$$

We also name it *multivariate distribution*. We explained before the independence of a pair of events. Using the cumulative distribution function, we can now define the independence between random variables.

Definition 13. A finite set of n random variables $\{X_1, \dots, X_n\}$ is mutually independent if and only if, for any sequence $\{x_1, \dots, x_n\}$, the events $\{X_1 \leq x_1\}, \dots, \{X_n \leq x_n\}$ are mutually independent. Equivalently, this finite set is mutually independent if and only if:

$$F_{X_1,\dots,X_n}(x_1, \dots, x_n) = F_{X_1}(x_1) \dots F_{X_n}(x_n), \qquad \text{for all } x_1, \dots, x_n.$$

We can also extend the notion of expectation to a random vector. Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random vector and assume that $E[X_i]$ exists for all $i \in \{1, \dots, n\}$. The expectation of $\mathbf{X}$ is defined as the vector containing the expectations of each individual random vector, that is:

$$E[\mathbf{X}] = \begin{bmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{bmatrix}.$$

To generalize the variance of a random variable, we have to build the following matrix.

Definition 14. Let $\mathbf{X} = (X_1, \dots, X_n)$ ve a random vector. Then, the *covariance matrix* of $\mathbf{X}$ is defined as:

$$\Sigma = \mathrm{Cov}(\mathbf{X}) = E[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^T] = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix},$$

where $\sigma_{ij} = Cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] = \sigma_{ji}$.

It can also happen that, given a *random vector*, we would like to know the probability distribution of a few of its components. That is called the *marginal distribution*.

Definition 15 (Marginal Distribution)**.** Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector. The marginal distribution of a subset of $\mathbf{X}$ is the probability distribution of the variables contained in the subset.

In the simple case of having two random variables, e.g. $X = (X_1, X_2)$, then the marginal distribution of X_1 is:

$$P(x) = \int_{X_2} P(x_1, x_2) dx_2.$$

2

DISTRIBUTIONS

We have introduced the concepts of *random variable*, *random vector* and its *probability distribution*. Now, given two distributions, in the following chapters we will like to see how different they are from each other. In order to compare them, we enunciate the definition of the Kullback-Leibler divergence.

Definition 16 (Kullback-Leibler Divergence)**.** Let P and Q be probability distributions over the same probability space Ω. Then, the Kullback-Leibler divergence is defined as:

$$D_{KL}(P \mid\mid Q) = E_P \left[\log \frac{P(x)}{Q(x)} \right].$$

It is defined if, and only if P is *absolutely continuous with respect to* Q, that is , if $P(A) = 0$ for any A subset of Ω where $Q(A) = 0$. There are some properties of this definition that must be stated. The first one is the following proposition:

Proposition 1. *If P, Q are two probability distributions over the same probability space, then $D_{KL}(P|Q) \geq 0$.*

Proof. Firstly, note that if $a \in \mathbb{R}^+$, then $\log\ a \leq a - 1$. Then:

$$\begin{aligned} -D_{KL}(P \mid\mid Q) &= -E_P \left[\log \frac{P(x)}{Q(x)} \right] \\ &= E_P \left[\log \frac{Q(x)}{P(x)} \right] \\ &\leq E_P \left[\left(\frac{Q(x)}{P(x)} - 1 \right) \right] \\ &= \int P(x) \frac{Q(x)}{P(x)} dx - 1 \\ &= 0. \end{aligned}$$

So we have obtained that $-D_{KL}(P \mid\mid Q) \leq 0$, which implies that $D_{KL}(P \mid\mid Q) \geq 0$. $\square$

As a corollary of this proposition, we can affirm that $D_{KL}(P \mid\mid Q)$ equals zero if and only if $P = Q$ almost everywhere. We will also remark the discrete case, as it will be used later. Let P, Q be discrete probability distributions defined on the same probability space Ω. Then,

$$D_{KL}(P \mid\mid Q) = \sum_{x \in \Omega} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

2.1 EXAMPLES OF DISTRIBUTIONS

Some examples of common distributions will now be presented. They will be used further in this document.

Bernoulli

Imagine that you want to model the possible outcomes of an experiment with two possibilites: sucess or failure. Imagine also that you already know that in your experiment there is a probability p of achieving success. That is the intuitive idea of a Bernoulli distribution. We can define it more formally as it follows:

The Bernoulli distribution is a discrete probability distribution of a random variable that takes two values, $\{0, 1\}$, with probabilities p and $q = 1 - p$, respectively. We will say that our distribution is a $Bern(p)$.

If k is a possible outcome, we can define the probability mass function f of a Bernoulli distribution as:

$$f(k, p) = \begin{cases} p & \text{if } k = 1, \\ 1 - p & \text{if} k = 0 \end{cases}$$

Using the expression of the mean for discrete random variables, we obtain that $E[X] = p$ and

$$\text{Var}[X] = E[X^2] - E[X]^2 = E[X] - E[X]^2 = p - p^2 = p(1 - p) = pq.$$

As a note, this is just a particular case of the *Binomial distribution* with $n = 1$.

Gaussian Distribution

The Gaussian (or normal) distribution is used to represent real-valued random variables whose distributions are not known. Its importancy relies in the fact that, using the *central limit theorem*, we can assume that the average of many samples of a random variable with finite mean and variance is a random variable whose distribution converges to a normal distribution.

Definition 17. We say that the real valued random variable X follows a *normal distribution* of parameters $\mu, \sigma \in \mathbb{R}$ if, and only if, its probability density function exists and it is determined by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

where μ is the mean and σ is its standard deviation. We denote this normal distribution as $X \sim \mathcal{N}(\mu, \sigma)$

The particular case where $\mu = 0$ and $\sigma = 1$ is widely used in statistics. In this case, the density function is simpler:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

As a remarkable property of these distributions is that, if $f : \mathbb{R} \to \mathbb{R}$ is a real valued function defined as $f(x) = ax + b$, then $f(X) \sim \mathcal{N}(\mu + b, a^2\sigma)$.

In the same way that we extended random variables to random vectors, we can extend the normal distribution to a multivariate random distribution.

Definition 18. We say that a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ follows a multivariate normal distributions of parameters $\mu \in \mathbb{R}^n$, $\Sigma \in \mathcal{M}_N(\mathbb{R})$ if, and only if, its probability density function is:

$$f(x) = \frac{1}{\sqrt{det(2\pi\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}.$$

It is denoted $X \sim \mathcal{N}(\mu, \Sigma)$. In this case, σ is the mean vector of the distribution and Σ denotes de covariance matrix.

2.2 PARAMETRIC MODELING

In the following chapters, we will be trying to estimate density functions in a dataset. To do this we will be using *parametric models*. We say that a *parametric model*, $p_\theta(x)$, is a family of density functions that can be described using a finite numbers of parameters θ. We can get to the concept of *log-likelyhood* now.

Definition 19. The *likelyhood* $\mathcal{L}(\theta|x)$ of a parameter set θ is a function that measures how plausible is θ, given an observed point x in the dataset $\mathcal{D}$. It is defined as the value of the density function parametrized by θ at x. That is:

$$\mathcal{L}(\theta|x) = p_\theta(x).$$

In a whole dataset $\mathcal{D}$ consisting in independent observations, we would can write:

$$\mathcal{L}(\theta|X) = \prod_{x \in D} p_\theta(x).$$

This can be computationally hard to work with, so it is often used the *log-likelyhood* instead.

Definition 20. Let $\mathcal{D}$ be a dataset of independent observatoins and θ a set of parameters. Then, we define the log-likelyhood $\updownarrow$ as the sum of the logarithms of the evaluations of p_θ in each x in the dataset. That is:

$$\updownarrow(\theta|X) = \sum_{x \in \mathcal{D}} \log p_\theta(x).$$

The optimal for our purposes would be to find the optimal value $\cap\theta$ that maximizes the likelyhood of observing the dataset $\mathcal{D}$. We get to the following definition:

Definition 21. We say that *maximum likelyhood estimation* is the method used

Part II

INFORMATION THEORY

Information theory is the base for all the following work. In this part, *Mutual Information* will be explained and then, bounds for this function will be given.

3

MUTUAL INFORMATION

Obtaining good representations of data is one of the most important tasks in Machine Learning. Recently, it has been discovered that maximizing *Mutual Information* between two elements in our data can give us good representations for our data. We will go through the basic concepts first.

3.1 ENTROPY

The *mutual information* concept is based on the *Shannon entropy*, which we will introduce first, along with some basic properties of it. The Shannon entropy is a way of measuring the uncertainty in a random variable. Given an event $\mathcal{A} \in \Omega$, P a probability measure and $P[\mathcal{A}]$ the probability of $\mathcal{A}$, we can affirm that

$$\log \frac{1}{P[\mathcal{A}]}$$

describes *how surprising is that $\mathcal{A}$ occurs*. For instance, if $P[\mathcal{A}] = 1$, then the last expression is zero, which means that it is not a surprise that $\mathcal{A}$ occurred. With this motivation, we get to the following definition.

Definition 22. Let X be a discrete random variable with image $\mathcal{X}$. The *Shannon entropy*, or simply *entropy* $H(X)$ of X is defined as:

$$H(X) = E_X\left[\log \frac{1}{P_X(X)}\right] = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)}$$

The *entropy* can trivially be expressed as:

$$H(X) = -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x)$$

There are some properties of the *entropy* that must be remarked.

Proposition 2. *Let X be a random variable with image $\mathcal{X}$. If $|\mathcal{X}|$ is the cardinal of $\mathcal{X}$, then*

$$0 \leq H(X) \leq \log(|\mathcal{X}|).$$

Proof. Since $\log y$ is concave on $\mathbb{R}^+$, by Jensen's inequality, see Appendix 1, Prop. 4, we obtain:

$$H(X) = -\sum_{x\in\mathcal{X}} P_X(x)\log P_X(x) \le \log\left(\sum_{x\in\mathcal{X}} 1\right) = \log(|\mathcal{X}|).$$

For the lower bound we see that, since $P_X(x) \in [0,1]$ for all $x \in \mathcal{X}$ then $\log P_X(x) \le 0 \ \forall x \in \mathcal{X}$. Hence , $-P_X(x)\log P_X(x) \ge 0$ for all $x \in X$, so $H(X) \ge 0$. □

We can also see that the equality on the left holds if , and only if , exists x in X such that its probability is exactly one, that is $P_X(x) = 1$. The right equality holds if and only if , for all $x \in \mathcal{X}$, its probability is $P_X(x) = \frac{1}{|X|}$.

Conditional entropy

We have already said that entropy measures how surprising is that an event occurs. Usually, we will be looking at two random variables and it would be interesting to see how surprising is that one of them, say X, occurred, if we already know that Y occurred. This leads us to the definition of *conditional entropy*. Let us see a simpler case first:

Let A be an event, and X a random variable. The conditional probability $P_{X|A}$ defines the entropy of X conditioned to A:

$$H(X|A) = \sum_{x\in\mathcal{X}} P_{X|A}(x)\log\frac{1}{P_{X|A}(x)}$$

If Y is another random variable and $\mathcal{Y}$ is its image, intuitively we can sum the conditional entropy of an event with all the events in $\mathcal{Y}$, and this way we obtain the conditional entropy of X given Y.

Definition 23 (Conditional Entropy)**.** Let X, Y be random variables with images $\mathcal{X}, \mathcal{Y}$. The *conditional entropy* $H(X|Y)$ is defined as:

$$\begin{aligned} H(X|Y) &:= \sum_{y\in\mathcal{Y}} P_{\mathcal{Y}}(y) H(X|Y=y) \\ &= \sum_{y\in\mathcal{Y}} P_{\mathcal{Y}}(y) \sum_{x\in\mathcal{X}} P_{X|Y}(x|y)\log\frac{1}{P_{X|Y}(x|y)} \\ &= \sum_{x\in X, y\in\mathcal{Y}} P_{XY}(x,y)\log\frac{P_Y(y)}{P_{XY}(x,y)}. \end{aligned}$$

The interpretation of the conditional entropy is simple: the uncertainty in X when Y is given. Since we know about an event that has occurred (Y), intuitively the conditional entropy , or the uncertainty of X occurring given that Y has occurred, will be lesser than the entropy of X, since we already have some information about what is happening. We can prove this:

Proposition 3. *Let X, Y be random variables with images $\mathcal{X}, \mathcal{Y}$. Then:*

$$0 \le H(X|Y) \le H(X).$$

Proof. The inequality on the left was proved on Proposition **??** 2. The characterization of when $H(X|Y) = 0$ was also mentioned after it. Let us look at the inequality on the right. Note that restricting to the (x, y) where $P_{XY}(x, y) > 0$ and using the definition of the conditional probability we have:

$$H(X|Y) = \sum_{y \in \mathcal{Y}} P_Y(y) \sum_{x \in \mathcal{X}} P_{X|Y}(x|y) \log \frac{1}{P_{X|Y}(x|y)}$$
$$= \sum_{x,y} P_Y(y) P_{X|Y}(x,y) \log \frac{P_Y(y)}{P_{XY}(x,y)} = \sum_{x,y} P_{XY}(x,y) \log \frac{P_Y(y)}{P_{XY}(x,y)}$$

and

$$H(X) = \sum_x P_X(x) \log \frac{1}{P_X(x)} = \sum_{x,y} P_{XY}(x,y) \log \frac{1}{P_X(x)}$$

hence,

$$H(X|Y) - H(X) = \sum_{x,y} P_{XY}(x,y) \left(\log \frac{P_Y(y)}{P_{XY}(x,y)} - \log \frac{1}{P_X(x)} \right) = \sum_{x,y} P_{XY} \log \frac{P_Y(y) P_X(x)}{P_{XY}(x,y)} \quad (1)$$

so, using Jensen's Inequality , we obtain:

$$\sum_{x,y} P_{XY} \log \frac{P_Y(y) P_X(x)}{P_{XY}(x,y)} \leq \log \left(\sum_{x,y} \frac{\cancel{P_{XY}(x,y)} \; P_Y(y) P_X(x)}{\cancel{P_{XY}(x,y)}} \right)$$
$$= \log \left(\left(\sum_x P_X(x) \right) \left(\sum_y P_Y(y) \right) \right) = \log 1 = 0,$$

and this leads us to:

$$H(X|Y) - H(X) \leq 0 \implies H(X|Y) \leq H(X)$$

as we wanted. □

It must be noted that, on the development of $H(X|Y) - H(X)$, in the first inequality, equality holds if and only if $P_{XY}(x, y) = P_X(x) P_Y(y)$ for all (x, y) with $P_{XY}(x, y) > 0$, as it is said in Jensen's inequality. For the second inequality, equality holds if and only if $P_{XY}(x, y) = 0$, which implies $P_X(x) P_Y(y) = 0$ for any $x \in \mathcal{X}$, $y \in \mathcal{Y}$. It follows that $H(X|Y) = H(X)$ if and only if $P_{XY}(x, y) = P_X(x) P_Y(y)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$

3.2 MUTUAL INFORMATION

Using the entropy of a random variable we can directly state the definition of *Mutual Information* as it follows:

Definition 24 (Mutual Information)**.** Let X, Z be random variables. The *Mutual Information (MI)* $I(X, Z)$ between X and Z is expressed as the difference between the entropy of X and the conditional entropy of X and Z, that is:

$$I(X, Z) := H(X) - H(X|Z)$$

Since the entropy of the random variable $H(X)$ explains the uncertainty of X occurring, the intuitive idea of the *MI* is to determine the decrease of uncertainty of X occurring when we already know that Z has occurred. We also have to note that, using the definition of the *entropy* and the expression obtained in Eq. 1, we can rewrite the *MI* as it follows:

$$\begin{aligned} I(X,Z) &= \sum_{x\in\mathcal{X}} P_X(x)\log\frac{1}{P(x)} - \sum_{x\in\mathcal{X},z\in\mathcal{Z}} P_{XZ}(x,z)\log\frac{P_Z(x)}{P_{XZ}(x,z)} \\ &= \sum_{x,z} P_{XZ}\log\frac{P_Z(z)P_X(x)}{P_{XZ}(x,z)} = D_{KL}(P_{XZ} \mid\mid P_X P_Z) \end{aligned}$$

and we have obtained an expression of the mutual information using the *Kullback-Leibler* divergence. This provides with the following immediate consequences:

(i) Mutual information is non-negative , that is : $I(X,Z) \geq 0$.

(ii) If X, Z are random variables, then its mutual information equals zero if, and only if, they are independent. This is trivial because if $D_{KL}(P_{XZ} \mid\mid P_X P_Z) = 0$, then $P_{XZ} = P_X P_Z$ almost everywhere so X and Z are independent.

(iii) Since $P_{XZ} = P_{ZX}$ and $P_X P_Z = P_Z P_X$, mutual information is symmetric. That is: $I(X,Z) = I(Z,X)$.

Part III

REPRESENTATION LEARNING

Clarification of this part

4

CONTEXT

Before continuing presenting the mathematical notions of the topics that are treated in this work, it is interesting to present what we are pursuing with this work.

Machine Learning is the part of computer science that studies *algorithms* that improve automatically through experience from examples. These algorithms help computer to discover how to perform tasks without being explicitly programmed to do them. For the computers to learn, it is mandatory that a finite set of data (or dataset) $\mathcal{D}$ is available.

Depending on how the data (*or signal*) is given to the computer, the machine learning approaches can be divided in three broad categories:

1. *Supervised learning*. In this category each point $x_i \in \mathcal{D}$ in the dataset is *labeled*: each example is related to a tag $y_i \in Y$ that gives information about x. The goal in this case is to find a function $g : D \to Y$.
2. *Unsupervised learning*. In ths case, the data is *unlabeled*, so the approach is completely different. Usually, the goal here is to discover hidden patterns in data or to learn features from it.
3. *Reinforcement learning*. This is the area concerned with how intelligent agents take decisions in an an specific environment in order to obtain the best reward in their objective.

MOTIVATION

In the learning process, machine learning models can not directly give labels to input examples. Before, they must create a *representation* that contains the data's key qualities. Here is where *representation learning* is born.

Representation learning is a set of techniques that allows a system to discover the representations needed for feature detection or classification. In contrast to manual feature engineering, feature learning allows a machine to learn the features and to use them to perform a task.

Feature learning can be supervised or unsupervised. In supervised feature learning, representations are learned using labeled data. Examples of this kind of feature learning are supervised neural networks and multilayer perceptron. In unsupervised learning, the features are learned using unlabeled data. There are many examples of this, such as independent component analysis (ICP) and autoencoders. In this work, we will be working with unsupervised feature learning.

The performance of machine learning methods is heavily dependent on the choice of data features Bengio *et al.* (2014). This is why most of the current effort in machine learning focuses on designing preprocessing and data transformation that lead to good quality representations. A representation will be of good quality when its features produce good results at running the models.

Data's features that are invariant through time are very useful for machine learning models. In Wiskott & Sejnowski (2002), *slow features* are presented. Slow features are defined as features of a signal (which can be the input of a model) that vary slowly during time. These kind of features are the most interesting ones when creating representations, since they give an abstract view of the original data.

Let us give an example: In computer vision, the value of the pixels in an image can vary fastly. For instance, if we have a zebra on a video and the zebra is moving from one side of the image to the other, due to the black stripes of this animal, the pixels will fastly change from black to white and viceversa, so value of pixels is probably not a good feature to choose as an slow feature. However, there will always be a zebra on the image, so the feature that indicates that there is a zebra on the image will stay positive throughout all the video, so we can say that this is a slow feature.

5

GENERATIVE MODELS

From now on, let $\mathcal{D}$ be any kind of observed data. This will always be a finite subset of samples taken from a probability distribution p_{data}. There are models that, given $\mathcal{D}$, try to approximate the probability distribution that lies underneath it. These are called *generative models (G.M.)*.

Generative models can give parametric and non parametric approximations to the distribution p_{data}. In our case, we will focus on parametric approximations where the model search for the parameters that minimize the *distance* between the model distribution and the data distribution.

We can express our problem more formally as it follows. Let θ be a generative model within a model family $\mathcal{M}$. The goal of generative models is the following optimization:

$$\min_{\theta \in \mathcal{M}} d(p_{data}, p_\theta),$$

where d stands for the distance between the distributions. We can use, for instance, $K - L$ divergence.

Generative models have many useful applications. We can however remark the characteristics that we would like our generative model to be able to do. Those are:

- Estimate the density function: given a datapoint , $x \in D$, estimate the probability of that point $p_\theta(x)$,
- Generate new samples from the model distribution $x \sim p_\theta(x)$,
- Learn useful features of the datapoints.

Retaking the example of the zebras, if make our generative model learn about images of zebras, we will expect our $p_\theta(x)$ to be high for zebra's images. We will also expect the model to generate new images of this animal and to learn different features of the animal, such as their big size in comparison with cats.

5.1 AUTOREGRESSIVE MODELS

A very first definition of *Autoregressive models (AR)* would be the following one: *autoregressive models are feed-forward models that predict future values using past values*. Let us go deeper into this concept and explain how it behaves.

Again, let $\mathcal{D}$ be a set of $n-$dimensional datapoints x. We can assume that $x \in \{0,1\}^n$ for simplicity, without losing generality. If we choose any $x \in \mathcal{D}$, using the chain rule of probability, we obtain

$$p(x) = \prod_{i=1}^{n} p(x_i|x_1,\ldots,x_{i-1}) = \prod_{i=1}^{n} p(x_i|\boldsymbol{x}_{<i}) \qquad \text{where} \qquad \boldsymbol{x}_{<i} = [x_1,\ldots,x_{i-1}]$$

We see using that expression how, fixing an order of the variables $x_1,\ldots,x_n$, the distribution for the $i-th$ random variable depends on all the preceeding values in the particular chosen order.

It is known that given a set of discrete and mutually dependent random variables, they can be displayed in a table of conditional probabilities. If K_i is the number of states that each random variable can take then $\prod K_i$ is the number of cells that the table will have. If we represent $p(x_i|\boldsymbol{x}_{<i})$ for every i in tabular form, we can represent any possible distribution over n random variables.

This, however, will cause an exponential growth on the complexity of the representation, because in our case we would need to specify 2^{n-1} possibilities for each case. In terms of neural networks, since each column must sum 1 because we are working with probabilities, we have $2^{n-1}-1$ parameters for this conditional, and the tabular representation becomes impractical for our network to learn.

In autoregressive generative models, the conditionals are specified as we have mentioned before: parametrized functions with a fixed numbers of parameters. More precisely, we assume the conditional distributions to be Bernoulli random variables and learn a function p_{θ_i} that maps these random variables to the mean of the distribution. Mathematically, we have to find

$$p_{\theta_i}(x_i|\boldsymbol{x}_{<i}) = Bern(f_i(x_1,\ldots,x_{i-1})),$$

where θ_i is the set of parameters that specify the mean function $f_i : \{0,1\}^{i-i} \to [0,1]$.

The number of parameters is the reduced to $\sum_{i=1}^{n}|\theta_i|$, and then we can not represent all possible distributions as we could when using the tabular form of the conditional probabilities. We are now setting the limit of its expressiveness because we are setting the conditional distributions $p_{\theta_i}(x_i|\boldsymbol{x}_{<i})$ to be *Bernoulli* random variables.

Let us see a very simple case first in order to understand it better and then we will generalize it. Let σ be a sigmoid non linear function and $\theta_i = \{\alpha_0^{(i)},\alpha_1^{(i)},\ldots,\alpha_{i-1}^{(i)}\}$ the parameters of the mean function. Then, we can define our function f_i as the application of the non linear function to the sum of the first parameter $\alpha_0^{(i)}$ with the product of each parameter $\alpha_j^{(i)}$ with its random variable x_j, with $j = 1,2,\ldots,i-1$. That is:

$$f_i(x_1,;x_{i-1}) = \sigma(\alpha_0^{(i)} + \alpha_1^{(i)}x_i + \cdots + \alpha_{i-1}^{(i)}x_{i-1}).$$

In this case, the number of parameters would be $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$, so using *Big O* notation, we would be in the case of $O(n^2)$. We will state now a more general and useful case, giving a more interesting parametrization for the mean function: *multi layer perceptrons*[1](MLP).

1 Multi layer perceptrons are feed-forward neural networks with at least 3 layers: input, hidden and output layers; each one using an activation function.

For this example we will consider the most simple MLP: the one with one hidden layer. Let $h_i = \sigma(\boldsymbol{A}_i \boldsymbol{x}_{<i} + c_i)$ be the hidden layer activation function. Remember that $h_i \in \mathbb{R}^d$. Let $\theta_i = \{\boldsymbol{A}_i \in \mathbb{R}^{d\times(i-1)},\ c_i \in \mathbb{R}^d,\ \alpha^{(i)} \in \mathbb{R}^d,\ b_i \in \mathbb{R}\}$ the set of parameters for the mean function f_i, that we define as:

$$f_i(\boldsymbol{x}_{<i}) = \sigma(\alpha^{(i)} h_i + b_i)$$

In this case, the number of parameters will be $O(n^2 d)$.

Part IV

A

Appendices

This appendix will be used to set forth some theoretical results that might not always be relevant but are needed to understand some details during this thesis. Not all of them will be proven.

Proposition 4 (Jensen's Inequality). *Let $f : \mathcal{D} \to \mathbb{R}$ be a concave function and $n \in \mathbb{N}$. For any $p_1, \ldots, p_n \in \mathbb{R}_0^+$ with $\sum p_i = 1$ and any $x_1, \ldots, x_n \in \mathbb{D}$, it holds that:*

$$\sum_{i=1}^{n} p_i f(x_i) \leq f(\sum_{i=1}^{n} p_i x_i)$$

Furthermore, if f is strictly *concave and $p_i \geq 0$ for all $i = 1, \ldots, n$,then the equality holds if and only if $x_1 = \cdots = x_n$*

BIBLIOGRAPHY

Bengio, Yoshua, Courville, Aaron, & Vincent, Pascal. 2014. Representation Learning: A Review and New Perspectives. *arXiv:1206.5538 [cs]*, Apr. arXiv: 1206.5538.

Löwe, Sindy, O'Connor, Peter, & Veeling, Bastiaan. 2019. Putting an End to End-to-End: Gradient-Isolated Learning of Representations. *Pages 3039–3051 of: Advances in Neural Information Processing Systems*.

Wiskott, Laurenz, & Sejnowski, Terrence J. 2002. Slow Feature Analysis: Unsupervised Learning of Invariances. *Neural Computation*, **14**(4), 715–770.