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, Sunday 6th December, 2020

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.

1.0.1 *Basic notions*

Definition 1. Let Ω be a non empty sample space. $\mathscr{A}$ is a $\sigma-$algebra over Ω if its a family of subsets of Ω that verify:

- $\emptyset \in \mathscr{A}$
- 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.

An *event* A is a subset of the finite sample space Ω. We define the probability of the subset as the sum of the probabilities of each of its points. That is:

$$\mathcal{P}(A) = \sum_{a\in A} \mathcal{P}(a)$$

Now, if B is another event, we will use the notation $P(A, B)$ to refer 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 contiditional probability we obtain that:

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

We also see fron the definition that

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

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

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

$\square$

Random variables will now be introduced. Their first propertty is that they are measurable functions. Those 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 $g \neq 0$) 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, μ) 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 $B \in (E, \mu)$, 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 *random variable* (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. We will refer to a *Random Variable* as a *R.V.*.

1.0.2 *Expectation of a random variable*

Definition 7. The *distribution function* F_X of a real-valued random variable X is:

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

And now, it will be introduced one of the most important concepts when talking about a *R.V.*, its expectation.

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:

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

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:

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

The *expectation* $E[X]$ of a *random variable* is a linear operation. That is, if $\mathcal{Y}$ 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:

$$E[X] = \sum_{x \in \mathcal{X}} x \dot{P}_X(x)$$

where x_i is each possible outcome of the experiment, and p_i the probability of the outcome x_i. The expression given in the definition 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*. Using *the mean* and we get to 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

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

Thanks to the linearity of the *expectation* of a random variable, it is easy to see that

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

2

DISTRIBUTIONS

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. Good features of our data will lead us to easier and more accurate training on *Artificial Neural Networks (ANNs)* and, thus, better results on experiments.

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* its a way of measuring the uncertainty in a random variable. Given an event $\mathcal{A} \in \Omega$, $\mathcal{P}$ a probability measure and $\mathcal{P}[\mathcal{A}]$ the probability of $\mathcal{A}$, we can affirm that

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

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.

3.1.1 *Discrete case*

Definition 11. Let X be a discrete random variable with image $\mathcal{X}$. The *Shannon entropy*, or simply *entropy* , $H(X)$ of X is defined as:

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

The *entropy* can trivially be expressed as:

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

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

Proposition 1. *Let X be a random variable with image $\mathcal{X}$. then*

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

Proof. Since $\log\ y$ is concave on $\mathbb{R}^+$, by Jensen's inequality , see 3,:

$$H(X) = -\sum_{x \in X} \mathcal{P}_X(x) \log \mathcal{P}_X(x) \leq \log(\sum_{x \in \mathcal{X}} 1) = \log(|\mathcal{X}|)$$

For the lower bound, it is easy to see that, since $\mathcal{P}_X(x) \in [0,1] \quad \forall x \in \mathcal{X}$, and, hence , $\log \mathcal{P}_X(x) \leq 0 \ \ \forall x \in \mathcal{X}$. The product of both is negative, so we have a sum of negative terms that is changed its sign afterwards, so it is always $H(X) \geq 0$. □

We can also see that the equality on the left holds if and only if $\exists x \in \mathcal{X} \ : P_X(x) = 1$. The right equality holds if and only if , forall $x \in \mathcal{X}$, $P_X(x) = \frac{1}{|X|}$

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, ocurred, if we already know that Y ocurred. This leads us to the definition of *conditional entropy*. Lets see a simpler case first:

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

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

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

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

The interpretation of the *Conditional Entropy* is simple: the uncertaincy in X when Y is given. Since we know about an event that has occurred (Y), intuitively the conditional entropy , or the uncertaincy 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 2. *Let X, Y be random variables with images $\mathcal{X}, \mathcal{Y}$. Then:*

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

Proof. The inequality on the left was proved on Proposition 1. The characterization of when $H(X|Y) = 0$ was also mentioned after it. Let's 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:

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

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,

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

so, using Jensen's Inequality , we obtain:

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

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 $(x, y) \in \mathcal{X} \times \mathcal{Y}$

Part III

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 3 (Jensen's Inequality). *Let $f : \mathcal{D} \to \mathbb{R}$ be a concave function and $n \in \mathbb{N}$. For any $p_1, \dots, p_n \in \mathbb{R}_0^+$ with $\sum p_i = 1$ and any $x_1, \dots, x_n \in \mathbb{D}$, 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 $i = 1, \dots, n$,then the equality holds if and only if $x_1 = \dots = x_n$*

BIBLIOGRAPHY

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*.