Introduction, review of probability and Bayes
Lesson No. 1

Samuel F. Chenatti

1 Probability

This chapter reviews the fundamentals aspects and tools of probability theory.

Since the text book does a very brief review of the basics of probability theory, we use Sheldon-Ross book [1] as a more complete reference.

For this reason the notation may diverge from the text book in some examples.

1.1 Experiments and events

In probability theory we define all the the possible **outcomes** of an unpredictable **experiment** as a set. In the discrete case the set is **finite or countable infinite**. This set is some times called the **sample space** from the **experiment**.

For the **experiment** of tossing a coing, we define the **sample space** as $S = \{H, T\}$ where $E = \{H\}$ is the **event** of the coin comming up heads and $E = \{T\}$ is the **event** of it coming up tail. Is this sense, an **event** of the **experiment** is any subset of the **sample space**.

If the experiment outcome lies in E, we say that the **event** ocurred. We define $P(E)$ as the of an event ocurring in an **experiment**.

Probability functions have the following main properties:

$$0 < p(E) < 1$$

$$P(S) = 1$$

The event $E = \{H, T\}$, for example, is the whole sample space, and $P(E) = 1$. For $E = \{H\}$, $P(E) = 1/2$ for an unbiased coin.

1.2 Discrete Random Variables

Be S the finite set of outcomes from the unpredictable **experiment**. Defining X as a **discrete random variable** of **experiment** implies that X is determined by the unpredictable **outcomes** of the experiment.

As an example, be the **experiment** of rolling a dice two times. Then the **sample space** is the finite set $S = \{\{1,1\},\{1,2\},\{1,3\},\{1,4\},\{1,5\},\{1,6\},\{2,1\},\{2,2\},...\}$ representing all the 36 possible **events** that can result from two dice rolls.

For an unbiased dice, all events have equal probability of happening, or $P(E) = 1/6, \forall E \subset S$.

Being X the random variable that denotes the sum of the two unbiased dice rolls. Then $X \in (X) = \{2,3,4,5,6,7,8,9,10,11,12\}$.

$$P\{X = 2\} = P\{\{1,1\}\} = 1/36$$

$$P\{X = 3\} = P\{\{1,2\},\{2,1\}\} = 2/36$$

$$P\{X = 4\} = P\{\{1,3\},\{3,1\},\{2,2\}\} = 3/36$$

And so on.

1.3 Probability of a union of two envents

For a single unbiased dice roll experiment, be $E = \{1,3,4\}$ and $F = \{1,2\}$ two events where $P(E) = 3/6$ and $P(F) = 3/6$ and we want to derive a formula for $P(E) + P(F)$. Since the resulting union of these two events is $\{1,2,3,4\}$, just summing up the odds would result in accounting the event $\{1\}$ two times. So we define:

$$P(E) + P(F) = P(E \cup F) + P(E \cap F) \rightarrow P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

Note that for mutually exclusive events (ex: $F' = \{2\}$) the intersection is the empty set, and then we have:

$$P(E \cup F') = P(E) + P(F')$$

1.4 Conditional probability

For the roll of two unbiased dices, be E the event of the sum of the dices being 6 and F the event of the first dice being a four. Than the probability of the sum of the dices is 6 **given that** the first dice is a four is denoted by:

$$P(E|F)$$

If the event F occurs, E can only occurs if the outcome is both a point in F and E (implying it is a point in $E \cap F$). Since F ocurred, it now becomes our sample space and then we have:

$$P(E|F) = P(E \cap F)/P(F)$$

1.5 Joint Probabilities

From the conditional probability, we define the joint probability of the joint event E and F as follows:

$$P(E) \cap P(F) = P(EF) = P(E|F)P(F)$$

Which is also called the **product rule**.

1.6 Marginal probabilities

The probability of a pedestrian being hit by a car given the semaphore light color can be modeled by two random variables: $E = \{Beinghit, Notbeinghit\}$ and $F = \{red, green, yellow\}$. The **marginal probability** of being hit is then defined as follows:

$$P(\{Beinghit\}) = P(\{Beinghit\}|\{red\}) + P(\{Beinghit\}|\{green\}) + P(\{Beinghit\}|\{yellow\})$$

Or, in a general form, we have the **sum rule**:

$$P(A) = \sum_b P(A \cup B) = \sum_b P(A|B = b)P(B = b)$$

1.7 Bayes' rule

The Bayes rule is defined as follows:

$$p(X = x|Y = y) = \frac{p(Y = y|X = x)p(X = x)}{\sum_{x'} p(X = x')p(Y = y|X = x'))}$$

1.8 Independence and conditional independence

We say that two events E and F are **unconditionally independent** or **marginally independent** if the joint probability is the product of the two marginals:

$$P(EF) = P(E)P(F)$$

Impliying that $P(E|F) = P(E)$, which means that the knowledge about F ocurring does not affect the probability of event E.

The events E and F are **conditionally independents** given an event G if the following equation holds:

$$P(EF|G) = P(E|G)P(F|G)$$

An ilustrated example of conditional independence can be found on the respective Wikipedia article.

1.9 Continuous random variables

Being X a **continuous random variable**, its set of possible values is uncountable since it is defined in a continuous space. For X, there is a nonnegative function $f(x)$ defined for all real $x \in (-inf, inf)$ called **probability density function** or **pdf** for short. A pdf function the property that for any set B of real numbers

$$P\{X \in B\} = \int_B f(x)dx$$

And, as defined for the discrete random variables,

$$P(X) \in (-inf, inf) = \int_{-inf}^{inf} f(x)dx = 1$$

Letting $B = a, b$ we then have

$$P\{a \leq X \leq b\} = \int_a^b f(x)dx$$

And for $a = b$ we have the following relation:

$$P\{a \leq X \leq b\} = \int_a^b f(x)dx = 0$$

Which implies that the probability of X assuming a particular value in the continuous space is 0. Most of the time we are concerned about finding the probaiblity of X belonging to a specific interval.

The **cumulative distribution function** is defined as follows:

$$F(a) = P\{X \in (-inf, a)\} = \int_a f(x)dx$$

$$\rightarrow \frac{dF(a)}{dx} = f(a)$$

2 Quantiles

Being F a monotonically increasing function, we can obtain the inverse function F^{-1}.

Remember that F maps the probability of the continuous random variable X assuming a value less or equal to x

$$F(x) = P(X \leq x) = p, 0 \leq p \leq 1$$

So the intuition is that $F^{-1}(p)$ tells us which value of x would make $F(x) = p$. We have the following notable quantiles:

- $F^{-1}(0.5)$ is the **median** of the distribution
- $F^{-1}(0.25)$ is the **lower quantile** of the distribution
- $F^{-1}(0.75)$ is the **upper quantile** of the distribution

3 Mean and variance

We can define the **expected value** or **mean** (μ) of a random variable X as follows:

- $\mathbb{E}[X] \triangleq \sum_{x \in \mathcal{X}} x p(x)$ in the discrete case
- $\mathbb{E}[X] \triangleq \int_{\mathcal{X}} x p(x) dx$ in the continuous case

Note that in both cases we define the expected value with respect to a specific probability density function or probability density function.

The **variance** (ρ^2) is defined as a measure of dispersion around the **mean**:

$$var[X] \triangleq \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - \mu^2$$
$$\rightarrow \mathbb{E}[X^2] = \mu^2 + \rho^2$$

The **standard deviation** is also a metric of dispersion, and is defined as follows:

$$std[X] \triangleq \sqrt{var[X]}$$

Different from variance, the standard deviation is defined in the same unit as X.

4 Monte Carlo approximation

We can use Monte Carlo to approximate the expected valur of any function of a random variable by computing the arithmetic mean of the function over sampled data:

$$\mathbb{E}[f(X)] \approx \frac{1}{S} \sum_{s=1}^{S} f(x_s)$$

It is intuitive from the above equation that the accuracy of Monte Carlo approximation increases with sample size.

As a side note, a notable application of Monte Carlo Sampling is in the field of Reinforcement Learning where we want to estimate the expected return of a policy given a batch of sampled trajectories from a simulation. More information can be found on Sutton and Barto book [2]

5 Information Theory

5.1 Entropy

The **entropy** $\mathbb{H}$ of a random variable X is a measure of **uncertainty**. For a discrete random variable with K states, we have:

$$\mathbb{H} \triangleq -\sum_{k=1}^{K} p(X=k) log_2 p(X=k)$$

Since we use the log in base 2, the unit is called *bits*.

It follows from the above equation that a distribution with all of its mass concentrated in one state has minimum entropy.

5.2 KL Divergence

The $\mathbb{KL}$ divergence is a way of measuring the dissimilarity of two probability distributions p and q:

$$\mathbb{KL}(p||q) \triangleq \sum_{k=1}^{K} p_k log \frac{p_k}{q_k}$$

Notice that the $\mathbb{KL}$ divergence **is not symmetric**, but it is sometimes refered as a metric of distance between two distributions.

From the above equation we can derive the **cross entropy**:

$$\mathbb{H}(p,q) \triangleq -\sum_{k} p_k log q_k$$

5.3 Mutual information

Being X and Y two random variables, we can defined how much knowing about one variable tells us about the other one by observing how close the joit distribution $p(XY)$ is to $p(X)p(Y)$ (the intuition behind this relation is defined in the joint probabilities subsection):

$$\mathbb{I}(X;Y) \triangleq \mathbb{KL}(p(X,Y))||p(X)p(Y)) = \sum_{x}\sum_{y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$$

References

[1] S. M. Ross, *Introduction to Probability Models*, Sixth. San Diego, CA, USA: Academic Press, 1997.

[2] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, 1st. Cambridge, MA, USA: MIT Press, 1998, ISBN: 0262193981.