Machine Learning Lecture Notes

Predrag Radivojac

January 13, 2015

1 Random Variables

Until now we operated on relatively simple sample spaces and produced measure functions over sets of outcomes. In many situations, however, we would like to use probabilistic modeling on sets (e.g. a group of people) where elements can be associated with various descriptors. For example, a person may be associated with his/her age, height, citizenship, IQ, or marital status and we may be interested in events related to such descriptors. In other situations, we may be interested in transformations of sample spaces such as those corresponding to digitizing an analog signal from a microphone into a set of integers based on some set of voltage thresholds. The mechanism of a *random variable* facilitates addressing all such situations in a simple, rigorous and unified manner.

A random variable is a variable that, from the observer's point of view, takes values non-deterministically, with generally different preferences for different outcomes. Mathematically, however, it is defined as function that maps one sample space into another, with a few technical caveats we will introduce later. A selection of $\mathbb{R}$ as the co-domain for the mapping defines a real-valued random variable.

Let us motivate the need for random variables. Consider a probability space $(\Omega, \mathcal{F}, P)$, where Ω is a set of people and let us investigate the probability that a randomly selected person $\omega \in \Omega$ has a cold (we may assume we have a diagnostic method and tools at hand). We start by defining an event A as

$$A = \{\omega \in \Omega : Disease(\omega) = \text{cold}\}$$

and simply calculate the probability of this event. This is a perfectly legitimate approach, but it can be much simplified using the random variable mechanism. We first note that, technically, our diagnostic method corresponds to a function $Disease : \Omega \to \{\text{cold}, \text{not cold}\}$ that maps the sample space Ω to a new binary sample space $\Omega_{Disease} = \{\text{cold}, \text{not cold}\}$. Even more interestingly, our approach also maps the probability distribution P to a new probability distribution $P_{Disease}$ that is defined on $\Omega_{Disease}$. For example, we can calculate $P_{Disease}(\{\text{cold}\})$ from the probability distribution of the aforementioned event A; i.e. $P_{Disease}(\{\text{cold}\}) = P(A)$. This is a cluttered notation so we may wish to simplify it by using $P(Disease = \text{cold})$, where $Disease$ is a "random variable".

We will use capital letters $X, Y, \ldots$ to denote random variables (such as *Disease*) and lowercase letters $x, y, \ldots$ to indicate elements (such as "cold") of $\Omega_X, \Omega_Y \ldots$ Generally, we write such probabilities as $P(X = x)$, which is a notational relaxation from $P(\{\omega : X(\omega) = x\})$. Before we proceed to formally define random variables, we shall look at two illustrative examples.

Example 3. Consecutive tosses of a fair coin. Consider a process of three coin tosses and two random variables, X and Y, defined on the sample space. We define X as the number of heads in the first toss and Y as the number of heads over all three tosses. Our goal is to find the probability spaces that are created after the transformations.

First, we have

ω	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
$X(\omega)$	1	1	1	1	0	0	0	0
$Y(\omega)$	3	2	2	1	2	1	1	0

Let us only focus on variable Y. Clearly, $Y : \Omega \to \{0, 1, 2, 3\}$ but we also need to find $\mathcal{F}_Y$ and P_Y. To calculate P_Y, a simple approach is to find its pmf p_Y. For example, let us calculate $p_Y(2) = P_Y(\{2\})$ as

$$\begin{aligned} P_Y(\{2\}) &= P(Y = 2) \\ &= P(\{\omega : Y(\omega) = 2\}) \\ &= P(\{\text{HHT}, \text{HTH}, \text{THH}\}) \\ &= \frac{3}{8}, \end{aligned}$$

because of the uniform distribution in the original space $(\Omega, \mathcal{F}, P)$. In a similar way, we can calculate that $P(Y = 0) = P(Y = 3) = 1/8$, and that $P(Y = 1) = 3/8$. In this example, we took that $\mathcal{F} = \mathcal{P}(\Omega)$ and $\mathcal{F}_Y = \mathcal{P}(\Omega_Y)$. As a final note, we mention that all the randomness is defined in the original probability space $(\Omega, \mathcal{F}, P)$ and that the new probability space $(\Omega_Y, \mathcal{F}_Y, P_Y)$ simply inherits it through a deterministic transformation.

□

Example 4. Quantization. Consider $(\Omega, \mathcal{F}, P)$ where $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}(\Omega)$, and P is induced by a uniform probability density function. Define $X : \Omega \to \{0, 1\}$ as

$$X(\omega) = \begin{cases} 0 & \omega \leq 0.5 \\ 1 & \omega > 0.5 \end{cases}$$

and find the transformed probability space.

Technically, we have changed the sample space to $\Omega_X = \{0, 1\}$. For an event space $\mathcal{F}_X = \mathcal{P}(\Omega_X)$ we would like to understand the new probability distribution P_X. We have

$$\begin{aligned} p_X(0) &= P_X(\{0\}) \\ &= P(X = 0) \\ &= P(\{\omega : \omega \in [0, 0.5]\}) \\ &= \frac{1}{2} \end{aligned}$$

and

$$\begin{aligned} p_X(1) &= P_X(\{1\}) \\ &= P(X = 1) \\ &= P(\{\omega : \omega \in (0.5, 1]\}) \\ &= \frac{1}{2} \end{aligned}$$

From here we can easily see that $P_X(\{0, 1\}) = 1$ and $P_X(\emptyset) = 0$, and so P_X is indeed a probability distribution. Again, P_X is naturally defined using P. Thus, we have transformed the probability space $(\Omega, \mathcal{F}, P)$ into $(\Omega_X, \mathcal{F}_X, P_X)$.

□

1.1 Formal definition of random variable

We now formally define a random variable. Given a probability space $(\Omega, \mathcal{F}, P)$, a random variable X is a function $X : \Omega \to \Omega_X$ such that for every $A \in \mathcal{B}(\Omega_X)$ it holds that $\{\omega : X(\omega) \in A\} \in \mathcal{F}$. It follows that

$$P_X(A) = P(\{\omega : X(\omega) \in A\}).$$

It is important to mention that, by default, we defined the event space of a random variable to be the Borel field of Ω_X. This is convenient because a Borel field of a countable set Ω is its power set. Thus, we are working with the largest possible event spaces for both discrete and continuous random variables.

Consider now a *discrete random variable* X defined on $(\Omega, \mathcal{F}, P)$. As we can see from the previous examples, the probability distribution for X can be found as

$$\begin{aligned} p_X(x) &= P_X(\{x\}) \\ &= P(\{\omega : X(\omega) = x\}) \end{aligned}$$

for $\forall x \in \Omega_X$. The probability of an event A can be found as

$$\begin{aligned} P_X(A) &= P(\{\omega : X(\omega) \in A\}) \\ &= \sum_{x \in A} p_X(x) \end{aligned}$$

for $\forall A \subseteq \Omega_X$.

The case of *continuous random variables* is more complicated, but reduces to an approach that is similar to that of discrete random variables. Here we first define a *cumulative distribution function* (cdf) as

$$\begin{aligned} F_X(t) &= P_X\left(\{x : x \leq t\}\right) \\ &= P_X\left((-\infty, t]\right) \\ &= P\left(X \leq t\right) \\ &= P\left(\{\omega : X(\omega) \leq t\}\right), \end{aligned}$$

where $P\left(X \leq t\right)$, as before, presents a minor abuse of of notation. If the cumulative distribution function is differentiable, the probability density function of a continuous random variable is defined as

$$p_X\left(x\right) = \left.\frac{dF_X\left(t\right)}{dt}\right|_{t=x}.$$

Alternatively, if p_X exists, then

$$F_X\left(t\right) = \int_{-\infty}^{t} p_X\left(x\right) dx,$$

for each $t \in \mathbb{R}$. Our focus will be exclusively on random variables that have their probability density functions; however, for a more general view, we should always keep in mind "if one exists" when referring to pdfs.

The probability that a random variable will take a value from interval $(a, b]$ can now be calculated as

$$\begin{aligned} P_X((a, b]) &= P_X(a < X \leq b) \\ &= \int_a^b p_X\left(x\right) dx \\ &= F_X\left(b\right) - F_X\left(a\right), \end{aligned}$$

which follows from the properties of integration.

Suppose now that the random variable X transforms a probability space $(\Omega, \mathcal{F}, P)$ into $(\Omega_X, \mathcal{F}_X, P_X)$. To describe the resulting probability space, we commonly use probability mass and density functions inducing P_X. For example, if P_X is induced by a Gaussian distribution with parameters μ and σ^2, we use

$$X : N(\mu, \sigma^2)$$

or sometimes $X \sim N(\mu, \sigma^2)$. Both notations indicate that the probability density function for the random variable X is

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

This distribution in turn provides all the necessary information about the probability space; i.e. $\Omega_X = \mathbb{R}$, and $\mathcal{F}_X = \mathcal{B}(\mathbb{R})$.

A group of k random variables $\{X_i\}_{i=1}^k$ defined on the same probability space $(\Omega, \mathcal{F}, P)$ is called a *random vector* or a multivariate (multidimensional) random variable. We have already seen an example of a random vector provided by random variables (X, Y) in Example 3. A generalization of a random vector to infinite sets is referred to as a *random process*; i.e. $\{X_i : i \in \mathcal{T}\}$, where $\mathcal{T}$ is an index set usually interpreted as a set of time indices. In the case of discrete time indices (e.g. $\mathcal{T} = \mathbb{N}$) the random process is called a discrete-time random process; otherwise (e.g. $\mathcal{T} = \mathbb{R}$) it is called a continuous-time random process.

1.2 Joint and marginal distributions

Let us first look at two discrete random variables X and Y defined on the same probability space $(\Omega, \mathcal{F}, P)$. We define the *joint probability distribution* $p_{XY}(x, y)$ of X and Y as

$$\begin{aligned} p_{XY}(x, y) &= P(X = x, Y = y) \\ &= P(\{\omega : X(\omega) = x\} \cap \{\omega : Y(\omega) = y\}). \end{aligned}$$

We can extend this to a k-dimensional random variable $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ and define a multidimensional probability mass function as $p_{\boldsymbol{X}}(\boldsymbol{x})$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$ is a vector of values, such that each x_i is chosen from some Ω_{X_i}.

A *marginal distribution* is defined for a subset of $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ by summing or integrating over the remaining variables. A marginal distribution $p_{X_i}(x_i)$ is defined as

$$p_{X_i}(x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_k} p_{\boldsymbol{X}}(x_1, \ldots, x_k),$$

where the variable in the j-th sum takes values from Ω_{X_j}. The previous equation directly follows from Eq. (1) in the Probability Theory section and is also referred to as *sum rule*.

In the continuous case, we define a multidimensional cdf as

$$\begin{aligned} F_{\boldsymbol{X}}(\boldsymbol{t}) &= P_{\boldsymbol{X}}(\{\boldsymbol{x} : x_i \leq t_i, i = 1 \ldots k\}) \\ &= P(X_1 \leq t_1, X_2 \leq t_2 \ldots) \end{aligned}$$

and the probability density function, if it exists, is defined as

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \left. \frac{\partial^k}{\partial t_1 \cdots \partial t_k} F_{\boldsymbol{X}}(t_1, \ldots t_k) \right|_{\boldsymbol{t}=\boldsymbol{x}}.$$

The marginal density $p_{X_i}(x_i)$ is defined as

$$p_{X_i}(x_i) = \int_{x_1} \cdots \int_{x_{i-1}} \int_{x_{i+1}} \cdots \int_{x_k} p_{\boldsymbol{X}}(\boldsymbol{x})\, dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_k$$

If the sample space for each discrete random variable is seen as a countable subset of $\mathbb{R}$, then the probability space for any k-dimensional random variable $\boldsymbol{X}$ (discrete or continuous) can be defined as $\left(\mathbb{R}^k, \mathcal{B}(\mathbb{R})^k, P_{\boldsymbol{X}}\right)$.

Example 5. Three tosses of a fair coin (again). Consider two random variables from Example 3 and calculate their probability spaces, joint and marginal distributions.

A joint probability mass function $p(x, y) = P(X = x, Y = y)$ is shown below

		Y			
		0	1	2	3
X	0	$1/8$	$1/4$	$1/8$	0
	1	0	$1/8$	$1/4$	$1/8$

but let us step back for a moment and show how we can calculate it. Let us consider two sets $A = \{\,\text{HHH, HHT, HTH, HTT}\,\}$ and $B = \{\,\text{HHT, HTH, THH}\,\}$, corresponding to the events that the first toss was heads and that there were exactly two heads over the three tosses, respectively. Now, let us look at the probability of the intersection of A and B

$$\begin{aligned} P(A \cap B) &= P(\{\text{HHT}, \text{HTH}\}) \\ &= \frac{1}{4} \end{aligned}$$

We can represent the probability of the logical statement $X = 1 \wedge Y = 2$ as

$$\begin{aligned} p_{XY}(1, 2) &= P(X = 1, Y = 2) \\ &= P(A \cap B) \\ &= P(\{\text{HHT}, \text{HTH}\}) \\ &= \frac{1}{4}. \end{aligned}$$

The marginal probability distribution can be found in a straightforward way as

$$p_X(x) = \sum_{y \in \Omega_Y} p_{XY}(x, y),$$

where $\Omega_Y = \{0, 1, 2, 3\}$. Thus,

$$\begin{aligned} p_X(0) &= \sum_{y \in \Omega_Y} p_{XY}(0, y) \\ &= \frac{1}{2}. \end{aligned}$$

We note for the end that we needed $|\Omega_X| \cdot |\Omega_Y| - 1$ numbers (because the sum must equal 1) to fully describe the joint distribution $p_{XY}(x, y)$. Asymptotically, this corresponds to an exponential growth of the number of entries in the table with the number of random variables (k). For example, if $|\Omega_{X_i}| = 2$ for $\forall X_i$, there are $2^k - 1$ free elements in the joint probability distribution. Estimating such distributions from data is intractable and is one form of the *curse of dimensionality*.

□

1.3 Conditional distributions

The conditional probability distribution for two random variables X and Y, $p(y|x)$, is defined as

$$\begin{aligned} p_{Y|X}(y|x) &= P(Y = y|X = x) \\ &= P(\{\omega : Y(\omega) = y\} \mid \{\omega : X(\omega) = x\}) \\ &= \frac{P(\{\omega : X(\omega) = x\} \cap \{\omega : Y(\omega) = y\})}{P(\{\omega : X(\omega) = x\})} \\ &= \frac{P_{XY}(X = x, Y = y)}{P_X(X = x)} \\ &= \frac{p_{XY}(x, y)}{p_X(x)} \end{aligned}$$

or simply $p(x, y) = p(y|x) \cdot p(x)$. This is a direct consequence of the product rule from Eq. (2) in the Probability Theory section. From this expression, we can calculate the probability of an event A as

$$P_{Y|X}(Y \in A|X = x) = \begin{cases} \sum_{y \in A} p_{Y|X}(y|x) & Y : \text{discrete} \\ \\ \int_{y \in A} p_{Y|X}(y|x)dy & Y : \text{continuous} \end{cases}$$

The extension to more than two variables is straightforward. We can write

$$p(x_k|x_1, \ldots, x_{k-1}) = \frac{p(x_1, \ldots, x_k)}{p(x_1, \ldots, x_{k-1})}.$$

By a recursive application of the product rule, we obtain

$$p(x_1, \ldots, x_k) = p(x_1) \prod_{l=2}^{k} p(x_l|x_1, \ldots, x_{l-1}) \tag{1}$$

which is referred to as *chain rule*.

1.4 Independence of random variables

Two random variables are independent if their joint probability distribution can be expressed as

$$p_{XY}(x, y) = p_X(x) \cdot p_Y(y).$$

As before, k random variables are (collectively, mutually) independent if a joint probability distribution of any subset of variables can be expressed as a product of individual (marginal) probability distributions of its components.

Another, different, form of independence can be found even more frequently in probabilistic calculations. It represents independence between variables in the presence of some other random variable (evidence), e.g.

$$p_{XY|Z}(x, y|z) = p_{X|Z}(x|z) \cdot p_{Y|Z}(y|z)$$

and is referred to as *conditional independenc*e. Interestingly, the two forms of independence are unrelated; i.e. neither one implies the other. We show this in two simple examples from Figure 1.

1.5 Expectations and moments

Expectations of functions are defined as sums (or integrals) of function values weighted according to the probability distribution function. Given a probability space $(\Omega_X, \mathcal{B}(\Omega_X), P_X)$, we consider a function $f : \Omega_X \to \mathbb{C}$ and define its expectation function as

$$E_x[f(x)] = \begin{cases} \sum_{x \in \Omega_X} f(x) p_X(x) & X : \text{discrete} \\ \\ \int_{\Omega_X} f(x) p_X(x) dx & X : \text{continuous} \end{cases}$$

It can happen for continuous distributions that $E_x[f(x)] = \pm\infty$; in such cases we say that the expectation does not exist. For $f(x) = x$, we have a standard expectation $E[X] = E_x[x] = \sum x p_X(x)$, or the mean value of X. Using $f(x) = x^k$ results in the k-th moment, $f(x) = \log \frac{1}{p_X(x)}$ gives the well-known entropy function $H(X)$, or differential entropy for continuous random variables, and $f(x) = (x - E[X])^2$ provides the variance of a random variable X, denoted by $V[X]$. Interestingly, the probability of some event $A \subseteq \Omega_X$ can also be expressed in the form of expectation; i.e.

$$P_X(A) = E[I_A(x)],$$

where

$$I_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

is an indicator function. With this, it is possible to express the cumulative distribution function as $F_X(t) = E[I_{(-\infty,t]}(x)]$.

A. X and Y are independent, but not conditionally independent given Z

$P(X=1)$
a

X	$P(Y=1\|X)$
0	b
1	b

X	Y	$P(Z=1\|X,Y)$
0	0	c
0	1	$1-c$
1	0	$1-c$
1	1	c

$$P(Y=y|X=x) = P(Y=y)$$

for example,

$$P(Y=1|X=x) = b$$
$$P(Y=1) = b$$

$$P(Y=y|X=x, Z=z) \neq P(Y=y|Z=z)$$

for example,

$$P(Y=1|X=1, Z=1) = a^2bc(1-c-b(1-2c))$$
$$P(Y=1|Z=1) = d^{-1}b(1-c-a(1-2c))$$

B. X and Z are conditionally independent given Y, but not independent

$P(X=1)$
a

X	$P(Y=1\|X)$
0	b
1	c

X	Y	$P(Z=1\|X,Y)$
0	0	d
0	1	e
1	0	d
1	1	e

$$P(Z=z|X=x) \neq P(Z=z)$$

for example,

$$P(Z=1|X=1) = d + ce - cd$$
$$P(Z=1) = d + (e-d)(a(c-b)+b)$$

$$P(Z=z|X=x, Y=y) = P(Z=z|Y=y)$$

for example,

$$P(Z=1|X=x, Y=1) = d$$
$$P(Z=1|Y=1) = d$$

Figure 1: Independence vs. conditional independence using probability distributions involving three binary random variables. Probability distributions are presented using factorization $p(x,y,z) = p(x)p(y|x)p(z|x,y)$, where all constants $a, b, c, d, e \in [0,1]$. (A) Variables X and Y are independent, but not conditionally independent given Z. When $c = 0$, $Z = X \oplus Y$, where $\oplus$ is an "exclusive or" operator. (B) Variables X and Z are conditionally independent given Y, but are not independent.

$f(x)$	Symbol	Name
x	$E[X]$	Mean
$(x - E[X])^2$	$V[X]$	Variance
x^k	$E[X^k]$	k-th moment; $k \in \mathbb{N}$
$(x - E[X])^k$	$E[(x - E[X])^k]$	k-th central moment; $k \in \mathbb{N}$
e^{tx}	$M_X(t)$	Moment generating function
e^{itx}	$\varphi_X(t)$	Characteristic function
$\log \frac{1}{p_X(x)}$	$H(X)$	(Differential) entropy
$\log \frac{p_X(x)}{q(x)}$	$D(p_X\|\|q)$	Kullback-Leibler divergence
$\left(\frac{\partial}{\partial\theta} \log p_X(x\|\theta)\right)^2$	$\mathcal{I}(\theta)$	Fisher information

Table 1: Some important expectation functions $E_x[f(x)]$ for a random variable X described by its distribution $p_X(x)$. Function $q(x)$ in the definition of the Kullback-Leibler divergence is non-negative and must sum (integrate) to 1; i.e. it is a probability distribution itself. The Fisher information is defined for a family of probability distributions specified by a parameter θ. Note that the moment generating function may not exist for some distributions and all values of t; however, the characteristic function always exists, even when the density function does not.

Function $f(x)$ inside the expectation can also be complex-valued. For example, $\varphi_X(t) = E_x[e^{itx}]$, where i is the imaginary unit, defines the characteristic function of X. The characteristic function is closely related to the inverse Fourier transform of $p_X(x)$ and is useful in many forms of statistical inference. Several expectation functions are summarized in Table 1.

Given two random variables X and Y and a specific value x assigned to X, we define the conditional expectation as

$$E_y[f(y)|x] = \begin{cases} \sum_{y\in\Omega_Y} f(y) p_{Y|X}(y|x) & Y : \text{discrete} \\ \int_{\Omega_Y} f(y) p_{Y|X}(y|x) dy & Y : \text{continuous} \end{cases}$$

where $f : \Omega_Y \to \mathbb{C}$ is some function. Again, using $f(y) = y$ results in $E[Y|x] = \sum y p_{Y|X}(y|x)$ or $E[Y|x] = \int y p_{Y|X}(y|x) dy$. We shall see later that under some conditions $E[Y|x]$ is referred to as *regression function*. These types of integrals are often seen and evaluated in Bayesian statistics.

For two random variables X and Y we also define

$$E_{x,y}\left[f(x,y)\right] = \begin{cases} \sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} f(x,y)p_{XY}(x,y) & X,Y : \text{discrete} \\ \\ \int_{\Omega_X}\int_{\Omega_Y} f(x,y)p_{XY}(x,y)dxdy & X,Y : \text{continuous} \end{cases}$$

and note that we interchangeably use $E\left[XY\right]$ and $E\left[xy\right]$. Expectations can also be defined over a single variable

$$E_{x}\left[f(x,y)\right] = \begin{cases} \sum_{x\in\Omega_X} f(x,y)p_{X}(x) & X : \text{discrete} \\ \\ \int_{\Omega_X} f(x,y)p_{X}(x)dx & X : \text{continuous} \end{cases}$$

Note that $E_x\left[f(x,y)\right]$ is a function of y.

We define the covariance function as

$$\begin{aligned} \text{cov}\left(X,Y\right) &= E_{x,y}\left[\left(x - E\left[X\right]\right)\left(y - E\left[Y\right]\right)\right] \\ &= E_{x,y}\left[xy\right] - E_x\left[x\right]E_y\left[y\right] \\ &= E\left[XY\right] - E\left[X\right]E\left[Y\right], \end{aligned}$$

with $\text{cov}(X) = \text{cov}(X,X)$ being the variance of the random variable X. Similarly, we define a correlation function as

$$\text{corr}\left(X,Y\right) = \frac{\text{cov}\left(X,Y\right)}{\sqrt{\text{cov}\left(X\right)\cdot\text{cov}\left(Y\right)}},$$

which is simply a covariance function normalized by the product of standard deviations. Both covariance and correlation functions have wide applicability in statistics, machine learning, signal processing and many other disciplines. Several important expectations for two random variables are listed in Table 2.

Example 6. Three tosses of a fair coin (yet again). Consider two random variables from Examples 3 and 5, and calculate the expectation and variance for both X and Y. Then calculate $E\left[Y|X=0\right]$.

We start by calculating $E\left[X\right] = 0\cdot p_X(0) + 1\cdot p_X(1) = \frac{1}{2}$. Similarly,

$$\begin{aligned} E\left[Y\right] &= \sum_{y=0}^{3} y\cdot p_Y(y) \\ &= p_Y(1) + 2p_Y(2) + 3p_Y(3) \\ &= \frac{3}{2} \end{aligned}$$

The conditional expectation can be found as

$f(x,y)$	Symbol	Name
$(x - E[X])(y - E[Y])$	$\text{cov}(X,Y)$	Covariance
$\frac{(x-E[X])(y-E[Y])}{\sqrt{V[X]V[Y]}}$	$\text{corr}(X,Y)$	Correlation
$\log \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}$	$I(X;Y)$	Mutual information
$\log \frac{1}{p_{XY}(x,y)}$	$H(X,Y)$	Joint entropy
$\log \frac{1}{p_{X\|Y}(x\|y)}$	$H(X\|Y)$	Conditional entropy

Table 2: Some important expectation functions $E_{x,y}[f(x,y)]$ for two random variables, X and Y, described by their joint distribution $p_{XY}(x,y)$. Mutual information is sometimes referred to as *average mutual information*.

$$\begin{aligned} E[Y|X=0] &= \sum_{y=0}^{3} y \cdot p_{Y|X}(y|0) \\ &= p_{Y|X}(1|0) + 2p_{Y|X}(2|0) + 3p_{Y|X}(3|0) \\ &= 1 \end{aligned}$$

where $p_{Y|X}(y|x) = p_{XY}(x,y)/p_X(x)$.

□

In many situations we need to analyze more than two random variables. A simple two-dimensional summary of all pairwise covariance values involving k random variables $X_1, X_2, \ldots X_k$ is called the covariance matrix. More formally, the covariance matrix is defined as

$$\mathbf{\Sigma} = [\Sigma_{ij}]_{i,j=1}^{k}$$

where

$$\begin{aligned} \Sigma_{ij} &= \text{cov}(X_i, X_j) \\ &= E_{x_i,x_j}\left[(x_i - E[X_i])(x_j - E[X_j])\right]. \end{aligned}$$

Here, the diagonal elements of a $k \times k$ covariance matrix are individual variance values for each variable X_i and the off-diagonal elements are the covariance values between pairs of variables. The covariance matrix is symmetric. It is sometimes called a variance-covariance matrix.

1.5.1 Properties of expectations

Here we review, without proofs, some useful properties of expectations. For any two random variables X and Y, and constant $c \in \mathbb{R}$, it holds that:

1. $E[cX] = cE[X]$
2. $E[X + Y] = E[X] + E[Y]$
3. $V[c] = 0$
4. $V[X] \geq 0$
5. $V[cX] = c^2 V[X]$.

In addition, if X and Y are independent random variables, it holds that:

1. $E[XY] = E[X] \cdot E[Y]$
2. $V[X + Y] = V[X] + V[Y]$
3. $\text{cov}(X, Y) = 0$.

1.6 Mixtures of distributions

In previous sections we saw that random variables are often described using particular families of probability distributions. This approach can be generalized by considering mixtures of distributions, i.e. linear combinations of other probability distributions. As before, we shall only consider random variables that have their probability mass or density functions.

Given a set of m probability distributions, $\{p_i(x)\}_{i=1}^m$, a finite mixture distribution function, or *mixture model*, $p(x)$ is defined as

$$p(x) = \sum_{i=1}^{m} w_i p_i(x), \tag{2}$$

where $\boldsymbol{w} = (w_1, w_2, \ldots, w_m)$ is a set of non-negative real numbers such that $\sum_{i=1}^m w_i = 1$. We refer to $\boldsymbol{w}$ as mixing coefficients or, sometimes, as mixing probabilities. A linear combination with such coefficients is called a convex combination. It is straightforward to verify that a function defined in this manner is indeed a probability distribution.

Here we will briefly look into the basic expectation functions of the mixture distribution. Suppose $\{X_i\}_{i=1}^m$ is a set of m random variables described by their respective probability distribution functions $\{p_{X_i}(x)\}_{i=1}^m$. Suppose also that a random variable X is described by a mixture distribution with coefficients $\boldsymbol{w}$ and probability distributions $\{p_{X_i}(x)\}_{i=1}^m$. Then, assuming continuous random variables defined on $\mathbb{R}$, the expectation function is given as

$$
\begin{aligned}
E_x\left[f(x)\right] &= \int_{-\infty}^{+\infty} f(x) p_X(x) dx \\
&= \int_{-\infty}^{+\infty} f(x) \sum_{i=1}^{m} w_i p_{X_i}(x) dx \\
&= \sum_{i=1}^{m} w_i \int_{-\infty}^{+\infty} f(x) p_{X_i}(x) dx \\
&= \sum_{i=1}^{m} w_i E_{x_i}[f(x)].
\end{aligned}
$$

We can now apply this formula to obtain the mean, when $f(x) = x$ and the variance, when $f(x) = (x - E[X])^2$, of the random variable X as

$$
E[X] = \sum_{i=1}^{m} w_i E[X_i],
$$

and

$$
V[X] = \sum_{i=1}^{m} w_i V[X_i] + \sum_{i=1}^{m} w_i \left(E[X_i] - E[X]\right)^2,
$$

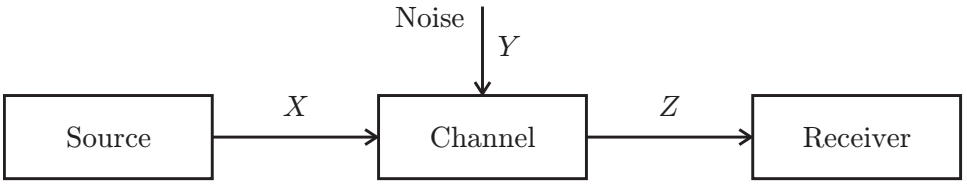
respectively. A mixture distribution can also be defined for countably and uncountably infinite numbers of components. Such distributions, however, are rare in practice.

Example 7. Signal communications. Consider transmission of a single binary digital signal (bit) over a noisy communication channel shown in Figure 2. The magnitude of the signal X emitted by the source is equally likely to be 0 or 1 Volt. The signal is sent over a transmission line (e.g. radio communication, optical fiber, magnetic tape) in which a zero-mean normally distributed noise component Y is added to X. Derive the probability distribution of the signal $Z = X + Y$ that enters the receiver.

We will consider a slightly more general situation where X : Bernoulli(α) and Y : Gaussian(μ, σ^2). To find $p_Z(z)$ we will use characteristic functions of random variables X, Y and Z, written as $\varphi_X(t) = E[e^{itx}]$, $\varphi_Y(t) = E[e^{ity}]$ and $\varphi_Z(t) = E[e^{itz}]$. Without derivation we write

$$
\begin{aligned}
\varphi_X(t) &= 1 - \alpha + \alpha e^{it} \\
\varphi_Y(t) &= e^{it\mu - \frac{\sigma^2 t^2}{2}}
\end{aligned}
$$

and subsequently



X: Bernoulli(α) $Z = X + Y$

Y: Gaussian(μ, σ^2)

Figure 2: A digital signal communication system with additive noise.

$$\begin{aligned}\varphi_Z(t) &= \varphi_{X+Y}(t) \\ &= \varphi_X(t) \cdot \varphi_Y(t) \\ &= \left(1 - \alpha + \alpha e^{it}\right) \cdot e^{it\mu - \frac{\sigma^2 t^2}{2}} \\ &= \alpha e^{it(\mu+1) - \frac{\sigma^2 t^2}{2}} + (1-\alpha) e^{it\mu - \frac{\sigma^2 t^2}{2}}.\end{aligned}$$

By performing integration on $\varphi_Z(t)$ we can easily verify that

$$p_Z(z) = \alpha \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(z-\mu-1)^2} + (1-\alpha) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(z-\mu)^2},$$

which is a mixture of two normal distributions $N(\mu+1, \sigma^2)$ and $N(\mu, \sigma^2)$ with coefficients $w_1 = \alpha$ and $w_2 = 1 - \alpha$, respectively. Observe that a convex combination of random variables $Z = w_1 X + w_2 Y$ does not imply $p_Z(z) = w_1 p_X(x) + w_2 p_Y(y)$.

□

1.7 Graphical representation of probability distributions

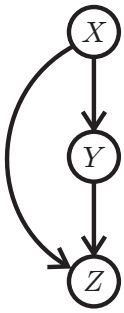
We saw earlier that a joint probability distribution can be *factorized* using the chain rule from Eq. (1). Such factorizations can be visualized using a directed graph representation, where nodes represent random variables and edges depict conditional dependence. For example,

$$p(x, y, z) = p(x)p(y|x)p(z|x, y)$$

is shown in Figure 3A.

Such visualization is especially convenient for showing conditional independence, which can be represented as missing edges in the graph. Figure 3B shows the the same factorization of $p(x, y, z)$ where variable Z is independent of X given Y.

A



$P(X=1)$
0.3

X	$P(Y=1\|X)$
0	0.5
1	0.9

X	Y	$P(Z=1\|X,Y)$
0	0	0.3
0	1	0.1
1	0	0.7
1	1	0.4

$$P(X=x, Y=y, Z=z) = P(X=x)P(Y=y|X=x)P(Z=z|X=x, Y=y)$$

B

$P(X=1)$
0.3

X	$P(Y=1\|X)$
0	0.5
1	0.9

Y	$P(Z=1\|Y)$
0	0.2
1	0.7

$$P(X=x, Y=y, Z=z) = P(X=x)P(Y=y|X=x)P(Z=z|Y=y)$$

Figure 3: Bayesian network: graphical representation of two joint probability distributions for three discrete (binary) random variables (X, Y, Z) using directed acyclic graphs. The probability mass function $p(x, y, z)$ is defined over $\{0, 1\}^3$. (A) Full factorization; (B) Factorization that shows and ensures conditional independence between Z and X, given Y. Each node is associated with a conditional probability distribution. In discrete cases, these conditional distributions are referred to as conditional probability tables.

Graphical representations of probability distributions using directed acyclic graphs, together with conditional probability distributions, are called *Bayesian networks*. They facilitate interpretation as well as effective statistical inference. Given a set of k random variables $\boldsymbol{X} = (X_1, \ldots, X_k)$, Bayesian networks factorize the joint probability distribution of $\boldsymbol{X}$ as

$$p(\boldsymbol{x}) = \prod_{i=1}^{k} p\left(x_i | \boldsymbol{x}_{\text{Parents}(X_i)}\right),$$

where Parents(X) denotes the immediate ancestors of node X in the graph. In Figure 3B, node Y is a parent of Z, but node X is not a parent of Z.

It is important to mention that there are multiple (how many?) ways of factorizing a distribution. For example, by reversing the order of variables $p(x, y, z)$ can be also factorized as

$$p(x, y, z) = p(z)p(y|z)p(x|y, z),$$

which has a different graphical representation and its own conditional probability distributions, yet the same joint probability distribution as the earlier factorization. Selecting a proper factorization and estimating the conditional probability distributions from data will be discussed in detail later.

Undirected graphs can also be used to factorize probability distributions. The main idea here is to decompose graphs into maximal cliques $\mathcal{C}$ (the smallest set of cliques that covers the graph) and express the distribution in the following form

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(\boldsymbol{x}_C),$$

where each $\psi_C(\boldsymbol{x}_C) \geq 0$ is called the clique potential function and

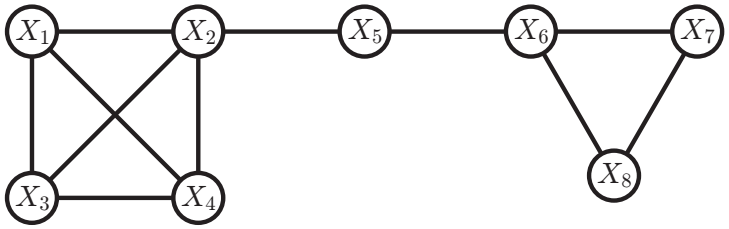
$$Z = \int_{\boldsymbol{x}} \prod_{C \in \mathcal{C}} \psi_C(\boldsymbol{x}_C) d\boldsymbol{x},$$

is called the partition function, used strictly for normalization purposes. In contrast to conditional probability distributions in directed acyclic graphs, the clique potentials usually do not have conditional probability interpretations and, thus, normalization is necessary. One example of a maximum clique decomposition is shown in Figure 4.

The potential functions are typically taken to be strictly positive, $\psi_C(\boldsymbol{x}_C) > 0$, and expressed as

$$\psi_C(\boldsymbol{x}_C) = \exp\left(-E(\boldsymbol{x}_C)\right),$$

where $E(\boldsymbol{x}_C)$ is a user-specified energy function on the clique of random variables $\boldsymbol{X}_C$. This leads to the probability distribution of the following form



$$\boldsymbol{X}_{C_1} = \{X_1, X_2, X_3, X_4\} \qquad \boldsymbol{X}_{C_3} = \{X_5, X_6\}$$
$$\boldsymbol{X}_{C_2} = \{X_2, X_5\} \qquad \boldsymbol{X}_{C_4} = \{X_6, X_7, X_8\}$$

Figure 4: Markov network: graphical representation of a probability distribution using maximum clique decomposition. Shown is a set of eight random variables with their interdependency structure and maximum clique decomposition (a clique is fully connected subgraph of a given graph). A decomposition into maximum cliques covers all vertices and edges in a graph with the minimum number of cliques. Here, the set of variables is decomposed into four maximal cliques $\mathcal{C} = \{C_1, C_2, C_3, C_4\}$.

$$p(\boldsymbol{x}) = \frac{1}{Z} \exp\left(\sum_{C \in \mathcal{C}} \log \psi_C(\boldsymbol{x}_C)\right).$$

As formulated, this probability distribution is called the Boltzmann distribution or the Gibbs distribution.

The energy function $E(\boldsymbol{x})$ must be lower for for values of $\boldsymbol{x}$ that are more likely. It also may involve parameters that are then estimated from the available training data. Of course, in a prediction problem, an undirected graph must be created to also involve the target variables, which were here considered to be a subset of $\boldsymbol{X}$.

Consider now any probability distribution over all possible configurations of the random vector $\boldsymbol{X}$ with its underlying graphical representation. If the following property

$$p\left(x_i | \boldsymbol{x}_{-X_i}\right) = p\left(x_i | \boldsymbol{x}_{N(X_i)}\right) \tag{3}$$

is satisfied, the probability distribution is referred to as *Markov network* or a *Markov random field*. In the equation above

$$\boldsymbol{X}_{-X_i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k)$$

and $N(X)$ is a set of random variables neighboring X in the graph; i.e. there exists an edge between X and every node in $N(X)$. The set of random variables in $N(X)$ is also called the Markov blanket of X.

It can be shown that every Gibbs distribution satisfies the property from Eq. (3) and, conversely, that for every probability distribution for which Eq. (3)

holds can be represented as a Gibbs distribution with some choice of parameters. This equivalence of Gibbs distributions and Markov networks was established by the Hammersley-Clifford theorem.