A Primer on Probability, Expectation, and Variance

Phil Schniter

January 26, 2022

1 Random variables

A *random variable* (RV) is a quantity that can take on many possible values. We never know what value the RV will take on, but we do know which values are possible and how often they occur.

A good analogy is a giant bag of marbles, where each marble has a number on it. We never know the particular number will appear when we draw a marble from the bag. These drawn values are called *realizations*, and the RV is the entire bag. The information contained by the RV includes the complete set of values written on the marbles, as well as the fraction of the bag that is taken by each of those values.

In this document, we will use capital letters for RVs (e.g., "A") and lowercase letters for deterministic variables (e.g., "a") in an attempt to be clear. But we can't do this throughout all of our lecture derivations, because that would conflict with many other notational conventions that we (and the machine learning community) have adopted.

When we write things out mathematically, we have to make a distinction between discrete RVs (i.e., those that take on values from a countable set) and continuous RVs (i.e., those that take on values from an uncountably infinite set).

1.1 Discrete random variables

A discrete RV A is often described using the *probability mass function* (pmf) $\{p_A[k]\}_{k=1}^K$, where integer k indexes over the K possible values in the set. For example, suppose that $A \in \{a^{(1)}, a^{(2)}, \ldots, a^{(K)}\}$. Then $p_A[k] = \Pr\{A = a^{(k)}\}$, i.e., the probability that A takes on the value $a^{(k)}$. The pmf must be non-negative and sum to one, i.e.,

$$p_A[k] \geq 0 \text{ for all } k, \text{ and } \sum_{k=1}^{K} p_A[k] = 1. \tag{1}$$

1.2 Continuous random variables

A continuous RV A is often described using the *probability density function* (pdf) $p_A(\cdot)$, which is defined for any real-valued input. The pdf must be non-negative and integrate to one, i.e.,

$$p_A(a) \geq 0 \text{ for all } a \in \mathbb{R}, \text{ and } 1 = \int_{-\infty}^{\infty} p_A(a)\,\mathrm{d}a. \tag{2}$$

Since it's a density, the pdf can be used to determine the probability of the event that A takes on values in a range, say $[a_1, a_2]$:

$$\Pr\{A \in [a_1, a_2]\} = \int_{a_1}^{a_2} p_A(a)\,\mathrm{d}a. \tag{3}$$

But notice that, when $p_A(a)$ is finite-valued for all a, there is zero probability that A takes on any particular value, like a_1. This can be seen by taking the limit $a_2 \to a_1$ in (3): the probability $\Pr\{A = a_1\} = \lim_{a_2 \to a_1} \Pr\{A \in [a_1, a_2]\}$ vanishes.

Note: it is possible to describe discrete a RV using a pdf as follows:

$$p_A(a) = \sum_{k=1}^{K} p_A[k]\delta(a - a^{(k)}), \tag{4}$$

where $\delta(\cdot)$ is the "point mass" or Dirac delta function. In words, $\delta(\cdot)$ is an infinitely tall, infinitely narrow, unit-area spike centered at the origin (a strange beast!). Mathematically,

$$\delta(0) = \infty, \quad \delta(a) = 0 \text{ for any } a \neq 0, \text{ and } \int_{-\infty}^{\infty} \delta(a)\,\mathrm{d}a = 1. \tag{5}$$

1.3 Joint random variables

Often we encounter expressions with multiple RVs that may be statistically related. We can describe this collection of RVs using a *joint pmf* (if they are all discrete) or a *joint pdf* more generally. In the case of two RVs A and B, we would have the joint pdf $p_{A,B}(\cdot, \cdot)$. Not surprisingly, $p_{A,B}$ must be non-negative and integrate to one, i.e.,

$$p_{A,B}(a, b) \geq 0 \text{ for all } a, b \in \mathbb{R}, \text{ and } 1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{A,B}(a, b)\,\mathrm{d}a\,\mathrm{d}b. \tag{6}$$

Also, $p_{A,B}(a, b) = p_{B,A}(b, a)$ for all $a, b \in \mathbb{R}$.

We say that A and B are *statistically independent* when

$$p_{A,B}(a, b) = p_A(a)p_B(b) \text{ for all } a, b \in \mathbb{R}. \tag{7}$$

As we will see below, that means you can't predict anything about A from B (or vice versa).

1.4 Conditional probability

One way to quantify the dependence among RVs is through *conditional probability*. Given two RVs A and B, the *conditional pdf of A given B* is written as $p_{A|B}(\cdot|\cdot)$. The quantity $p_{A|B}(a|b)$ describes the pdf of A evaluated at a given the event that $B = b$.

Example: Suppose that you model temperature as a continous random variable A and whether or not it is snowing as a binary random variable B (i.e., $B = 1$ means it is snowing). Then $p_{A|B}(a|1)$ would be the pdf of temperatures given that it is snowing. And $p_{B|A}(1|a)$ would be probability that it is snowing given that the temperature equals a.

The conditional and joint distributions can be connected as follows:

$$p_{A|B}(a|b)\, p_B(b) = p_{A,B}(a, b). \tag{8}$$

Also, the order of conditioning can be "swapped" using *Bayes rule*:

$$p_{A|B}(a|b) = \frac{p_{B|A}(b|a)p_A(a)}{p_B(b)}, \tag{9}$$

which follows from (8) and the equality $p_{A,B}(a,b) = p_{B,A}(b,a)$.

In the special case that A and B are statistically independent, (7) and (8) imply

$$p_{A|B}(a|b)\, p_B(b) = p_A(a)\, p_B(b) \tag{10}$$

$$\Leftrightarrow \qquad p_{A|B}(a|b) = p_A(a) \tag{11}$$

which says that the density of A is invariant to the value of B. Thus, you can't predict anything about A from B. By exchanging A and B in (8), you can similarly prove that $p_{B|A}(b|a) = p_B(b)$, i.e., you can't predict anything about B from A.

2 Expectation

The *expected value* $\mathrm{E}\{A\}$ of a RV A is the average or "mean" value of A. Mathematically, it has the following definition:

$$\text{discrete}: \quad \mathrm{E}\{A\} = \sum_{k=1}^{K} a^{(k)}\, p_A[k] \tag{12}$$

$$\text{continuous}: \quad \mathrm{E}\{A\} = \int_{-\infty}^{\infty} a\, p_A(a)\, \mathrm{d}a \tag{13}$$

It's important to understand the distinction between this statistical average of A and the *sample average* of a set of realizations $\{a_1, \ldots, a_n\}$ of A, which we write as

$$\overline{a} \triangleq \frac{1}{n} \sum_{i=1}^{n} a_i. \tag{14}$$

For example, when the statistical average is zero (i.e., $\mathrm{E}\{A\} = 0$), the sample average may still be non-zero. Concretely, if A is a zero-mean Gaussian random variable, then $\mathrm{E}\{A\} = 0$ but $\overline{a} \neq 0$ for any finite value of n. (You can easily check this in Python or Matlab.)

There are two key properties of expectation that we will use: For any deterministic constants c and d, deterministic functions $f(\cdot)$ and $g(\cdot)$, and independent RVs A and B, we have

1. Linearity: $\mathrm{E}\{c + d\, f(A)\} = c + d\, \mathrm{E}\{f(A)\}$
2. Independence: $\mathrm{E}\{f(A)g(B)\} = \mathrm{E}\{f(A)\}\ \mathrm{E}\{g(B)\}$

3 Variance

The *variance* $\mathrm{var}\{A\}$ of a RV AA is the mean squared deviation from the mean of A. Mathematically, it has the following definition:

$$\text{discrete}: \quad \mathrm{var}\{A\} = \mathrm{E}\left\{(A - \mathrm{E}\{A\})^2\right\} = \sum_{k=1}^{K} (a^{(k)} - \mathrm{E}\{A\})^2\, p_A[k] \tag{15}$$

$$\text{continuous}: \quad \mathrm{var}\{A\} = \mathrm{E}\left\{(A - \mathrm{E}\{A\})^2\right\} = \int_{-\infty}^{\infty} (a - \mathrm{E}\{A\})^2\, p_A(a)\, \mathrm{d}a \tag{16}$$

Note that the statistical variance is not equal to the *sample variance* of set of realizations $\{a_1, \ldots, a_n\}$ of A, which we write (using the sample mean $\overline{a}$) as

$$s_a^2 \triangleq \frac{1}{n} \sum_{i=1}^{n} (a_i - \overline{a})^2. \tag{17}$$

One very useful identity is

$$\mathrm{E}\{A^2\} = \mathrm{var}\{A\} + (\mathrm{E}\{A\})^2. \tag{18}$$

A short proof is now provided using the shorthand notation $\mu \triangleq \mathrm{E}\{A\}$:

$$\begin{aligned} \mathrm{E}\{A^2\} &= \mathrm{E}\{[(A-\mu)+\mu]^2\} && (19)\\ &= \mathrm{E}\{(A-\mu)^2 + 2(A-\mu)\mu + \mu^2\} && (20)\\ &= \underbrace{\mathrm{E}\{(A-\mu)^2\}}_{\mathrm{var}\{A\}} + 2\underbrace{(\mathrm{E}\{A\} - \mu)}_{=0}\mu + \underbrace{\mu^2}_{(\mathrm{E}\{A\})^2}. && (21) \end{aligned}$$

4 Conditional expectation and variance

Given several RVs, we sometimes want to describe the statistical average of one variable given fixed values of the rest. We can do this using *conditional expectation*. For example, given two RVs A and B, we define the *conditional mean of A given the event $B = b$* as follows:

$$\mathrm{E}\{A \,|\, B = b\} = \int_{-\infty}^{\infty} a \, p_{A|B}(a|b) \, \mathrm{d}a. \tag{22}$$

Note that this quantity is a deterministic function of the deterministic constant b, i.e.,

$$\mathrm{E}\{A \,|\, B = b\} = f(b) \tag{23}$$

for some function $f(\cdot)$. Sometimes we encounter the closely related quantity $\mathrm{E}\{A \,|\, B\}$. This is nothing more than the same function $f(\cdot)$ from (23) applied to the random variable B, i.e.,

$$\mathrm{E}\{A \,|\, B\} = f(B). \tag{24}$$

Thus $\mathrm{E}\{A \,|\, B\}$ is itself random, i.e., a random variable.

Similarly, we may want to know how far A deviates from its mean, on average, when $B = b$. We thus define the *conditional variance of A given the event $B = b$* as follows:

$$\begin{aligned} \mathrm{var}\{A \,|\, B = b\} &= \mathrm{E}\left\{(A - \mathrm{E}\{A \,|\, B\})^2 \;\middle|\; B = b\right\} && (25)\\ &= \int_{-\infty}^{\infty} (a - \mathrm{E}\{A \,|\, B = b\})^2 \, p_{A|B}(a|b) \, \mathrm{d}a. && (26) \end{aligned}$$

Note that $\mathrm{var}\{A \,|\, B = b\}$ is a deterministic function of b. We could also consider $\mathrm{var}\{A \,|\, B\}$, which is a function of the random variable B.

Next we consider some useful applications of conditional mean and variance.

4.1 Law of total expectation

Consider evaluating the quantity $\mathrm{E}\{f(A,B,C)\}$, which is the average value of a function $f(\cdot)$ involving the three RVs A, B, and C. In computing this quantity, we are "averaging out" the random contributions from A, B, and C. In some cases, when computing $\mathrm{E}\{f(A,B,C)\}$, it is more convenient to "average out" *one random variable at a time*. We can do this by *nesting the expectations*. For example, we could first remove the randomness due to C alone by holding A and B fixed via

$$g(A,B) \triangleq \mathrm{E}\{f(A,B,C) \,|\, A,B\}, \tag{27}$$

which gives a function g of RVs A and B. Next we could remove the randomness due to B by holding A fixed:

$$h(A) \triangleq \mathrm{E}\{g(A,B) \,|\, A\}. \tag{28}$$

And finally we could remove the randomness due to A:

$$\mathrm{E}\{h(A)\}. \tag{29}$$

The key point is that this latter quantity is the identical to what we would get from averaging A, B, C all at once, i.e.,

$$\begin{aligned}
\mathrm{E}\{f(A,B,C)\} &= \mathrm{E}\{h(A)\} && (30)\\
&= \mathrm{E}\{\ \mathrm{E}\{g(A,B) \,|\, A\}\ \} && (31)\\
&= \mathrm{E}\{\ \mathrm{E}\{\ \mathrm{E}\{f(A,B,C) \,|\, A,B\}\ \,|\, A\}\ \}. && (32)
\end{aligned}$$

This is known as the *law of total expectation*. It's typically stated in the 2-variable case, and without the function $f(\cdot)$, as follows:

$$\mathrm{E}\{A\} = \mathrm{E}\{\ \mathrm{E}\{A|B\}\ \}. \tag{33}$$

A short proof is now provided:

$$\begin{aligned}
\mathrm{E}\{\ \mathrm{E}\{A|B\}\ \} &= \int \left[\int a\, p_{A|B}(a|b)\, \mathrm{d}a \right] p_B(b)\, \mathrm{d}b && (34)\\
&= \int a \int \underbrace{p_{A|B}(a|b) p_B(b)}_{p_{A,B}(a,b)}\, \mathrm{d}b\, \mathrm{d}a && (35)\\
&= \int a\, p_A(a)\, \mathrm{d}a && (36)\\
&= \mathrm{E}\{A\}. && (37)
\end{aligned}$$

4.2 MMSE estimation

Another application of condition expectation is *mininum mean-squared error* (MMSE) estimation. Say that our goal is to predict a target variable y from vector of features $\boldsymbol{x} = [x_1, \ldots, x_d]^\top$ using a function $f(\boldsymbol{x})$. We'd like that, on average, the squared error $(y - f(\boldsymbol{x}))^2$ is minimized. We can formalize our notion of "average" using statistical expectation if we model the the target and features as random variables Y and $\boldsymbol{X} = [X_1, \ldots, X_d]^\top$.

In this statistical formulation, we want that $f(\cdot)$ minimizes the *mean-squared error*

$$\text{MSE} \triangleq \text{E}\left\{\big(Y - f(\boldsymbol{X})\big)^2\right\}. \tag{38}$$

We can derive the MMSE $f(\cdot)$ by first using the law of total expectation to write

$$\text{MSE} = \text{E}\{\ \underbrace{\text{E}\left\{\big(Y - f(\boldsymbol{X})\big)^2 \,\middle|\, \boldsymbol{X}\right\}}_{\text{MSE}(\boldsymbol{X})}\ \}. \tag{39}$$

If we can find the $f(\boldsymbol{X})$ that minimizes $\text{MSE}(\boldsymbol{X})$ for *any* $\boldsymbol{X}$, then this same function will minimize MSE when we average over $\boldsymbol{X}$ in (39). Notice that $\text{MSE}(\boldsymbol{X})$ can be written as

$$\begin{aligned}
\text{MSE}(\boldsymbol{X}) &= \text{E}\left\{\big(Y - f(\boldsymbol{X})\big)^2 \,\middle|\, \boldsymbol{X}\right\} && (40)\\
&= \text{E}\left\{Y^2 - 2Y f(\boldsymbol{X}) + f(\boldsymbol{X})^2 \,\middle|\, \boldsymbol{X}\right\} && (41)\\
&= \text{E}\left\{Y^2 \,|\, \boldsymbol{X}\right\} - 2\,\text{E}\{Y \,|\, \boldsymbol{X}\} f(\boldsymbol{X}) + f(\boldsymbol{X})^2. && (42)
\end{aligned}$$

The optimal $f(\boldsymbol{X})$, which we'll call $\widehat{f}(\boldsymbol{X})$, is simply find the value of $f(\boldsymbol{X})$ that sets the derivative of $\text{MSE}(\boldsymbol{X})$ to zero. In other words,

$$0 = \left.\frac{\partial \text{MSE}(\boldsymbol{X})}{\partial f(\boldsymbol{X})}\right|_{f(\boldsymbol{X})=\widehat{f}(\boldsymbol{X})} = -2\,\text{E}\{Y \,|\, \boldsymbol{X}\} + 2\widehat{f}(\boldsymbol{X}) \tag{43}$$

which implies

$$\widehat{f}(\boldsymbol{X}) = \text{E}\{Y \,|\, \boldsymbol{X}\}. \tag{44}$$

So, the MMSE estimator of Y from $\boldsymbol{X}$ is the conditional mean, $\widehat{f}(\boldsymbol{X}) = \text{E}\{Y \,|\, \boldsymbol{X}\}$.

Suppose that we would like to know the smallest possible value of $\text{MSE}(\boldsymbol{X})$, i.e., the value attained by the MMSE estimator $\widehat{f}(\boldsymbol{X})$. We can simply plug $\widehat{f}(\boldsymbol{X})$ from (44) into (40) and obtain

$$\begin{aligned}
\text{MSE}(\boldsymbol{X})\big|_{\text{min}} &= \text{E}\left\{\big(Y - \widehat{f}(\boldsymbol{X})\big)^2 \,\middle|\, \boldsymbol{X}\right\} && (45)\\
&= \text{E}\left\{\big(Y - \text{E}\{Y \,|\, \boldsymbol{X}\}\big)^2 \,\middle|\, \boldsymbol{X}\right\}, && (46)
\end{aligned}$$

which we recognize as the *conditional variance* of Y. So, the minimum value of $\text{MSE}(\boldsymbol{X})$ is the conditional variance, $\text{MSE}(\boldsymbol{X})\big|_{\text{min}} = \text{var}\{Y \,|\, \boldsymbol{X}\}$.