III. Statistical Estimation and Classification

Probability: An Extremely Concise Review

1. A scalar-valued random variable X is completely characterized by its distribution function

$$F_X(u) = \mathrm{P}\left(X \leq u\right).$$

This is also called the **cumulative distribution function** (cdf). $F_X(u)$ is monotonically increasing in u; it goes to one as $u \to \infty$ and goes to zero as $u \to -\infty$.

2. If F_X is differentiable, then we can also characterize X using its **probability density function** (pdf)

$$f_X(x) = \left.\frac{\mathrm{d}F_X(u)}{\mathrm{d}u}\right|_{u=x}$$

The density has the properties $f_X(x) \geq 0$ and

$$\int_{-\infty}^{\infty} f_X(x)\ \mathrm{d}x = 1.$$

Events of interest are subsets[1] of the real line — given such an event/subset $\mathcal{E}$, we can compute the probability of $\mathcal{E}$ occurring as

$$\mathrm{P}\left(\mathcal{E}\right) = \int_{x \in \mathcal{E}} f_X(x)\ \mathrm{d}x.$$

[1] Technically, it must be a subset of the real line that can be written as some combination of countable unions, countable intersections, and complements of intervals. You really have to know something about real analysis to construct a set that does not meet this criteria.

It is possible that a pdf exists even if F_X is not differentiable everywhere, for example:

$$F_X(u) = \begin{cases} 0, & u < 0, \\ u, & 0 \le u \le 1, \\ 1, & u \ge 1 \end{cases} \quad \text{has pdf} \quad f_X(x) = \begin{cases} 0, & x < 0, \\ 1, & 0 \le x \le 1, \\ 0, & x > 1. \end{cases}$$

3. The **expectation** of a function $g(X)$ of a random variable is

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) \ \mathrm{d}x.$$

This is the "average value" of $g(X)$ in that given a series of realizations $X = x_1, X = x_2, \ldots,$ of X,

$$\frac{1}{M} \sum_{m=1}^{M} g(x_m) \to \mathrm{E}[g(X)], \quad \text{as } M \to \infty.$$

This fact is known as the **(weak) law of large numbers**.

4. The **moment** of X of degree p is the expectation of the monomial $g(x) = x^p$. The zeroth moment is always 1:

$$\mathrm{E}[X^0] = \mathrm{E}[1] = \int_{-\infty}^{\infty} f_X(x) \ \mathrm{d}x = 1,$$

and the first moment is the **mean**:

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x \, f_X(x) \ \mathrm{d}x.$$

The **variance** is the second moment minus the mean squared:

$$\mathrm{var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \mathrm{E}[(X - \mathrm{E}[X])^2].$$

This is sometime referred to as the "variation around the mean". Aside from the zeroth moment, there is nothing that says that the integrals above must converge; it is easy to construct examples of well-defined random variables where $\mathrm{E}[X] = \infty$.

5. A pair of random variables (X, Y) are completely described by their **joint distribution function** (joint cdf)[2]

$$F_{X,Y}(u, v) = \mathrm{P}\left(X \leq u, Y \leq v\right).$$

Again, if $F_{X,Y}$ is continuously differentiable, (X, Y) is also characterize by the density

$$f_{X,Y}(x, y) = \left.\frac{\partial F_{X,Y}(u, v)}{\partial u \, \partial v}\right|_{(u,v)=(x,y)}.$$

In this case, events of interest correspond to regions in the plane $\mathbb{R}^2$, and the probability of an event occurring is the integral of the density over this region.

6. From the joint pdf $f_{X,Y}(x, y)$, we can recover the individual **marginal pdfs** for X and Y using

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \ \mathrm{d}y,$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \ \mathrm{d}x.$$

The pair of densities $f_X(x)$, $f_Y(y)$ tell us how X and Y behave individually, but not how they *interact*.

[2]For fixed $u, v \in \mathbb{R}$, the notation $\mathrm{P}\left(X \leq u, Y \leq v\right)$ should be read as "the probability that X is $\leq u$ and Y is $\leq v$.

7. If X and Y do interact in a meaningful way, then observing one of them affects the distribution of the other. If we observe $X = x$, then with this knowledge, the density for Y becomes

$$f_Y(y|X = x) = \frac{f_{X,Y}(x, y)}{f_X(x)}.$$

This is a density over y; it is easy to check that it is positive everywhere and that it integrates to one. $f_X(y|X = x)$ is called the **conditional density** for Y given $X = x$.

8. We call X and Y **independent** if observing X tells us nothing about Y (and vice versa). This means

$$f_Y(y|X = x) = f_Y(y), \quad \text{for all } x \in \mathbb{R},$$

and

$$f_X(x|Y = y) = f_X(x), \quad \text{for all } y \in \mathbb{R}.$$

(If one of the statements above is true, then the other follows automatically.) Equivalently, independence means that the joint pdf is *separable*:

$$f_{X,Y}(x, y) = f_X(x)\, f_Y(y).$$

9. We can always factor the joint pdf in two different ways:

$$f_X(x)f_Y(y|X = x) = f_{X,Y}(x, y) = f_Y(y)f_X(x|Y = y).$$

At this point, we should be comfortable enough with what is going on that we can use $f_Y(y|x)$ as short-hand notation for $f_Y(y|X = x)$. Then we can rewrite the above in its more common form as

$$f_X(x)f_Y(y|x) = f_{X,Y}(x, y) = f_Y(y)f_X(x|y).$$

This factorization also gives us a handy way to compute the marginals:

$$f_X(x) = \int_{-\infty}^{\infty} f_Y(y) f_X(x|y) \ \mathrm{d}y.$$

It also yields **Bayes' equation**

$$f_X(x|y) = \frac{f_Y(y|x) f_X(x)}{f_Y(y)},$$

which is a fundamental relation for statistical inference.

10. All of the above extends in the obvious way to more than two random variables. A **random vector**

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_D \end{bmatrix}$$

is completely characterized by the density $f_X(\boldsymbol{x}) = f_X(x_1, \ldots, x_D)$ on $\mathbb{R}^D$. In general, we can factor the joint pdf as

$$f_X(\boldsymbol{x}) = f_{X_1}(x_1)\, f_{X_2}(x_2|x_1)\, f_{X_3}(x_3|x_2, x_1) \cdots f_{X_D}(x_D|x_1, \ldots, x_{D-1}).$$

11. The pth moment of a random vector X that maps into $\mathbb{R}^D$ is the collection of expectations of all monomials of order p. The mean of a random vector is a vector of length D:

$$\mathrm{E}[X] = \begin{bmatrix} \mathrm{E}[X_1] \\ \vdots \\ \mathrm{E}[X_D] \end{bmatrix},$$

the second moment is the $D \times D$ matrix of all correlations between entries:

$$\mathrm{E}[XX^{\mathrm{T}}] = \begin{bmatrix} \mathrm{E}[X_1^2] & \mathrm{E}[X_1X_2] & \cdots & \mathrm{E}[X_1X_D] \\ \vdots & & \ddots & \vdots \\ \mathrm{E}[X_DX_1] & \cdots & & \mathrm{E}[X_D^2] \end{bmatrix},$$

the third moment is the $D \times D \times D$ tensor $\mathrm{E}[X \otimes X \otimes X]$, where

$$(\mathrm{E}[X \otimes X \otimes X])\,(i,j,k) = \mathrm{E}[X_iX_jX_k],$$

and so on. The **covariance** matrix contains all the pairs of second moments:

$$R_{i,j} = \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])].$$

If $\boldsymbol{\mu}_X = \mathrm{E}[X]$ is the mean vector, we can write the covariance matrix succinctly in terms of the second moment as

$$\boldsymbol{R} = \mathrm{E}[XX^{\mathrm{T}}] - \boldsymbol{\mu}_X\boldsymbol{\mu}_X^{\mathrm{T}}$$

12. Given independent observations $X = \boldsymbol{x}_1,\ X = \boldsymbol{x}_2, \ldots,\ X = \boldsymbol{x}_M$ of a random vector X with unknown (or partially known) distribution, a completely reasonable way to estimate the mean vector is using

$$\hat{\boldsymbol{\mu}} = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{x}_m.$$

If the mean $\boldsymbol{\mu}_X = \mathrm{E}[X]$ is known but the covariance is not, we can estimate the covariance using

$$\hat{\boldsymbol{R}} = \left(\frac{1}{M}\sum_{m=1}^{M} \boldsymbol{x}_m\boldsymbol{x}_m^{\mathrm{T}}\right) - \boldsymbol{\mu}_X\boldsymbol{\mu}_X^{\mathrm{T}}.$$

If both the mean and covariance are unknown, we first estimate the mean vector as above, then take

$$\hat{\boldsymbol{R}} = \left(\frac{1}{M-1} \sum_{m=1}^{M} \boldsymbol{x}_m \boldsymbol{x}_m^{\mathrm{T}} \right) - \hat{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}}^{\mathrm{T}}.$$

The difference in the scaling is to ensure that $\mathrm{E}[\hat{\boldsymbol{R}}] = \boldsymbol{R}$ in both cases.

Minimum Mean-Square Error Estimation

Now we will take our first look at estimating variables that are themselves random, subject to a known probability law.

We start our discussion with a very basic problem. Suppose Y is a scalar random variable with a known pdf $f_Y(y)$. Here is a fun game: you guess what Y is going to be, then I draw a realization of Y corresponding to its probability law, then we see how close you were with your guess.

What is your best guess?

Well, that of course depends on what exactly we mean by "best", i.e. what price I pay for being a certain amount off. But if we penalize the *mean-squared error*, we know exactly how to minimize it.

Let g be your guess. The error in your guess is of course random (since the realization of Y is random), and so is the squared-error $(Y-g)^2$. We want to choose g so that the mean of the squared error is as small as possible:

$$\underset{g}{\text{minimize}} \;\; \mathrm{E}[(Y-g)^2].$$

Expanding the squared error makes it clear how to do this:

$$\mathrm{E}[(Y-g)^2] = \mathrm{E}[Y^2] - 2g\,\mathrm{E}[Y] + g^2.$$

No matter what the first moment $\mathrm{E}[Y]$ and second moment $\mathrm{E}[Y^2]$ are (as long as they are finite), the expression above is a convex quadratic function in g, and hence is minimized when its first derivative (w.r.t. g) is zero, i.e. when

$$-2\,\mathrm{E}[Y] + 2g = 0 \quad \Rightarrow \hat{g} = \mathrm{E}[Y].$$

The squared error for this choice $\hat{g}$ is of course exactly the variance of Y.

The story gets more interesting (and relevant) when we have multiple random variables, some of which we observe, some of which we do not. Suppose that two random variables (Y, Z) have joint pdf $f_{Y,Z}(y, z)$. Suppose that a realization of (Y, Z) is drawn, and I get to observe Z. What have I learned about Y?

If Y and Z are independent, then the answer is of course nothing. But if they are not independent, then the marginal distribution of Y changes. In particular, before the random variables were drawn, the (marginal) pdf for Y was

$$f_Y(y) = \int f_{Y,Z}(y, z)\ \mathrm{d}z.$$

After we observe $Z = z$, we have

$$f_Y(y|Z=z) = \frac{f_{Y,Z}(y,z)}{f_Z(z)} = \frac{f_{Y,Z}(y,z)}{\int f_{Y,Z}(y,z)\ \mathrm{d}y}.$$

Y is still a random variable, but its distribution depends on the value z that was observed for Z.

Now, given that I have observed $Z = z$, what is the best guess for Y? If by "best" we mean that which minimizes the mean squared error, it is the conditional mean. That is, the minimizer of

$$\underset{g}{\text{minimize}} \;\; \mathrm{E}[(Y-g)^2|Z=z]$$

is

$$g = \mathrm{E}[Y|Z=z].$$

Notice that unlike before, g is not pre-determined, it depends on the outcome $Z = z$. We might denote

$$g(z) = \mathrm{E}[Y|Z=z].$$

In fact, since Z is a random variable, g is a priori also random, we might say

$$g(Z) = \mathrm{E}[Y|Z].$$

It is then fair to ask: what is the mean of $g(Z)$? We have[3]

$$\begin{aligned}\mathrm{E}[g(Z)] &= \mathrm{E}[\mathrm{E}[Y|Z]] \\ &= \int \mathrm{E}[Y|Z=z] f_Z(z) \; \mathrm{d}z \\ &= \mathrm{E}[Y]\end{aligned}$$

So on average, we are doing the same thing as if we didn't observe Z at all, but our MSE will in general be much better.

[3] That $\mathrm{E}[\mathrm{E}[Y|Z]] = \mathrm{E}[Y]$ is known as the law of *iterated expectation*. The inside E above is over Y while the outside one is over Z.