Gaussians

1 A Reminder in Probability

In this section we'll remind ourselves a bit of the basics of probability. Most of the definitions here will not be rigorous, and this is because it isn't the primary topic of... whatever this is.

1.1 Discrete Probability

The easiest place to begin is with discrete probability, however we will quickly depart into the realm of continuous probabilities, since that's where we'll dwell most of the time. In the discrete scheme, we assume that there is a space of "states" which we will call Ω for now. In this space of "states", let's assume that there are different discrete "states", i.e.:

$$\Omega = \{\omega_1, \omega_2, \omega_3, ...\}$$

The size of Ω may be finite or infinite, at this point it doesn't really matter.

The Probability Function

A *probability function* on Ω is any function $P : \Omega \to [0, 1]$ that satisfies the following[1]:

$$\sum_{\omega \in \Omega} P(\omega) = 1 \tag{1}$$

A very basic example would be a die; the state space would be $\Omega = \{1, 2, 3, 4, 5, 6\}$ and the probability function for $\omega \in \Omega$ is simply $P(\omega) = 1/6$, as you would expect.

Random Variables

This leads us to the next object of interest, the *random variable*. A random variable is a variable (that we define) that can take any value and is a placeholder for a state from Ω. This is confusing but intuitive if you have an example. Again, in the die example, we can define a random variable $X = \{1, 2, 3, 4, 5, 6\}$. Now, if we want to see how probable a singe value of X is, we can ask $P(X = \omega) =?$ which in this example will be like asking $P(\omega) =?$. This seems like a rather technical and unimportant point, but will help us later on. For now, we will abbreviate $P(X = \omega)$ into simply $P(\omega)$ - this is a typical abbreviation and is a bit more intuitive. A natural way to think about random variables is as a possible outcome of an experiment with the possible states Ω and the probability for each state $P(\omega)$. Of course, we can have multiple random variables at once. For two random variables X and Y, we will say that they are independent if and only if:

$$P(X = x \text{ and } Y = y) \triangleq P(x, y) = P(X = x) P(Y = y)$$

The way to think about this is "the outcome of experiment X does not effect the outcome of experiment Y". This is not always true, of course.

Expected Value, Variance and Covariance

A useful quantity to get acquainted with is the *expected value* of a random variable (or its *expectation*):

$$\mathbb{E}[X] \triangleq \sum_{\omega \in \Omega} P(X = \omega) \cdot \omega \tag{2}$$

This quantity tells us what value will be the mean of an infinite number of experiments with the random variable x. The expected value is a linear function, in other words:

$$\mathbb{E}[X + aY] = \mathbb{E}[X] = a\mathbb{E}[Y]$$

This will prove to be extremely helpful in the near future.

A few more definitions that we will need in the near future are *variance* and *standard deviation*, *covariance*:

[1] Note that, while not stated explicitly, $0 \leq P(\omega) \leq 1$ for any ω and there exists $\omega \in \Omega$ such that $P(\omega) > 0$. This is implied by the image of the function and eq. 1

1. Variance:

$$\mathrm{Var}[X] \triangleq \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}^2[X]$$

2. Standard deviation:

$$\sigma_X \triangleq \sqrt{\mathrm{Var}[X]}$$

3. Covariance:

$$\begin{aligned}\mathrm{Cov}[X,Y] &\triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathrm{Cov}[Y,X]\end{aligned}$$

In this case, if X is independent of Y, then $\mathrm{Cov}[Y,X] = 0$

Conditionality

There are cases where if we know the outcome of one experiment, this greatly changes the possible outcomes of another experiment. This is called conditionality, and specifically if we know the value of a random variable Y and want to check the probability for any outcome of the random variable X, we will denote this probability as:

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

Note that just from the definition we see that if X is independent from Y, then $P(x|y) = P(x)$. In general:

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

Finally, from this we can derive the following formulas:

1. $P(x,y) = P(x|y)\,P(y)$
2. Bayes' Theorem: $P(x|y) = \frac{P(y|x)P(x)}{P(y)}$
3. Law of total probability[2]: $P(x) = \sum_y P(x|y)\,P(y)$
4. Law of total expectation: $\mathbb{E}[X] = \sum_y \mathbb{E}[X|Y = y]\,P(y)$

And many other cases which will reveal themselves later on.

1.2 Continuous Variables

Up until now, all of the definitions we saw were for discrete variables, but when we try to extend this to continuous variables the definitions fall flat. For one, if we assume that there are an infinite number of "states" with non-zero probability, then $\sum_\omega P(\omega) > 1$. Otherwise, if we assume that they all have 0 probability, then $\sum_\omega P(\omega) = 0$. Clearly, then, this definition for probability is not adequate. Indeed, we will look at a different function to determine how likely values of our continuous variables are - the *probability density function* (PDF). In a moment we'll see how this relates to the definitions we saw before, for now we'll define a PDF as any function $f : X \to \mathbb{R}$ such that $\forall x\ f(x) \geq 0$ that satisfies:

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

This last condition is misleading as f can achieve values that are greater than 1 as long as the integration of all of its image is equal to 1. The reason this is called a density function is because we can now think of the probability that the function lies within a range of its values, i.e.:

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

So X is more likely to reside in areas where f is denser, but we still get $P(X = a) = \int_a^a f(x)\,dx = 0$, which makes sense in continuous variables. Note that f is not the probability function, only a middle man towards understanding the probability of finding the random variable in a range of values.

[2]This holds only if all possible states of Y partition Ω completely, but this will hold in most of the uses we'll see, so I've omitted it for now

We can now extend the definition of the expected value to continuous variables as well:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx \quad \text{"}\approx\text{"} \sum_x xP(x)$$

and can even extend the definition to functions of the random variable:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx$$

All of the other definitions we have seen above follow from this scheme and all of the properties still hold.

One more thing we should define is the *joint probability*; the PDF of two random variables x and y is defined as the function $f : \mathbb{R}^2 \to \mathbb{R}$ such that $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\, dxdy = 1$ and is non-negative. If x and y are independent, then

$$f(x,y) = f(x) f(y)$$

This is exactly as we saw with independent discrete variables, only now we have integrals everywhere (but we'll get used to them).

1.3 Using a Lot of Variables

There are some cases where we want to look at the probability of a lot of random variables at the same time. This could be, as a simple example, throwing n dice together and looking at the probability that each of them lands on a specific number. The way we defined the probability space above, we would need to look at the function $P(x_1, x_2, ..., x_n)$ which might be cumbersome to write. Instead, we can define a *random vector*. This is a vector that is comprised of random variables that is

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$$

where $x_1, x_2, ..., x_n$ are all random variables. In essence, this is nothing but a new way to notate what is actually going on. In this case, we will write

$$P(x_1, x_2, x_3, ..., x_n) \triangleq P(\boldsymbol{x})$$

so that we'll have a more concise way to denote the joint probability. This is of course the same for discrete and continuous variables.

Using this notation, we can define the *covariance matrix* Σ as follows:

$$\Sigma_{ij} = \mathrm{Cov}[x_i, x_j]$$

There are several things to note here, from this definition. First, the diagonal of Σ is actually the variance of each of the random variables, that is $\Sigma_{ii} = \mathrm{Var}[x_i]$. Second, notice that Σ must be symmetrical. Another important aspect is that if $\Sigma_{ij} = 0$ then x_i and x_j are independent. Finally, the covariance matrix is a *positive semi-definite* (PSD) matrix.

Positive Definitiveness of the Covariance Matrix

A PSD is matrix A has the following important quality:

$$\forall x \neq 0 \quad x^T A x \geq 0$$

This is as a result of the definition of the covariance. While this is always true, we will only think of the instances where the covariance matrix is a *positive definite* (PD) matrix, i.e. when:

$$\forall x \neq 0 \quad x^T A x > 0$$

The reason behind this is that if there exists a non-trivial vector x such that $x^T A x = 0$, this means that at least on of the values of the diagonal of A is equal to 0, i.e. there's a random variable with no variance - it is constant. This means that there is a random variable we can "drop", which will make our lives easier. Easier how? If a matrix is PD, it is invertible.

Further properties that PD matrices have are:

- All of the eigenvalues of the matrix are positive $\forall i \quad \lambda_i > 0$ (for PSD matrices they are non-negative)
- All PSD matrices have a root $A = RR^T$. If the matrix is PD, this root is unique. In both cases, the root can R is a lower triangular matrix (and the decomposition $A = RR^T$ is called the Cholesky decomposition)

1.4 Change of Variables

The final part in our probability chapter will be devoted to looking into the case where we want to change from one random vector to another. We can think of this as a sort of change of variable.

Recall that from the definition of our probability and PDF, they are defined on a specific random variable/set of random variables. Suppose that we have two (different) random vectors x and y with respective PDFs f_x and f_y, and that we have a function $g : X \to Y$ such that $y = g(x)$ that is one-to-one and differentiable. It would be ideal if we could somehow represent the probability of the random variable y using the same PDF as x. In this case, the following equation will be useful:

$$f_y(y) = f_x\left(g^{-1}(y)\right)\left|J\left(g^{-1}(y)\right)\right| \tag{3}$$

where $|J(g(y))|$ is the determinant of the Jacobian of the function g at the point y. This all seems rather difficult, but will help us later on in proving some basic properties in Gaussian distributions.

2 Gaussian Distribution

The Gaussian distribution is a distribution that widely used since it has certain favorable aspects that most distributions do not. For instance, the marginals of a Gaussian is a Gaussian, the product of i.i.d. Gaussians is a Gaussian, sampling is relatively easy, a linear transformation of a Gaussian is a Gaussian and on and on. The easiest place to start is in a single dimension.

2.1 1D Gaussian Distribution

The PDF of the Gaussian distribution is defined as:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{4}$$

For any $\sigma \geq 0$ and $\mu \in \mathbb{R}$. Since this will come up a lot, we will rewrite the above as:

$$p(x) \triangleq \mathcal{N}\left(x;\, \mu, \sigma^2\right) \Leftrightarrow x \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

So anytime you see something like $\mathcal{N}\left(\mu, \sigma^2\right)$, it will be a placeholder for eq. 4. The two most important properties of the Gaussian distribution are it's mean and standard deviation:

$$\mathbb{E}[x] = \mu$$
$$\sqrt{\mathrm{Var}[x]} = \sigma$$

and these are the only parameters needed to define the distribution. Further, we can easily see that the distribution is symmetrical and centered around μ and is uni-modal - it has only one mode. In this case that means that:

$$\mathbb{E}[x] = \arg\max_x p(x) = \mathrm{median}(x)$$

which are all useful properties for a distribution.

Add a visualization of the Gaussian distribution here.

A special case of the Gaussian distribution is called the *normal distribution*, which you have probably heard of. The normal distribution is

$$z \sim \mathcal{N}(0, 1)$$

and in fact every other Gaussian distribution is a linear transformation of the normal distribution (we will see this later on).

2.2 Multivariate Gaussian Distribution

The multivariate Gaussian distribution, also called Multivariate Normal Distribution (MVN) sometimes, is a generalization of the 1D case of the Gaussian distribution. This distribution is also completely defined by the mean and the variance, but since we are in the territory of multivariate variables (in other words, random vectors), we will need to be careful of how we define everything.

Let's start in the simplest generalization we can. Suppose that we have n different, **independent**, Gaussian random variables:

$$x_1, x_2, ..., x_n \sim \mathcal{N}(\mu_1, \sigma_1), \mathcal{N}(\mu_2, \sigma_2), ..., \mathcal{N}(\mu_n, \sigma_n)$$

We can, very simply, stack them together into a vector:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

But, how will this vector's distribution look? Let's check:

$$p(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(x_i-\mu_i)^2}{\sigma_i^2}\right]$$
$$= \frac{1}{(2\pi)^{n/2}} \frac{1}{\prod_i \sigma_i} \exp\left[-\frac{1}{2}\sum_i \frac{(x_i-\mu_i)^2}{\sigma_i^2}\right]$$

This doesn't seem to help us all that much... However, notice that if we define:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \sigma_n^2 \end{bmatrix} = I \begin{bmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \vdots \\ \sigma_n^2 \end{bmatrix}$$

as well as:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$$

then we get:

$$\sum_i \frac{(x_i-\mu_i)^2}{\sigma_i^2} = (x-\mu)^T \Sigma^{-1} (x-\mu)$$

This greatly decreases the amount we need to write. Also, notice how:

$$\prod_i \sigma_i = \sqrt{\prod_i \sigma_i^2} = \sqrt{|\Sigma|}$$

Suddenly, what we have before can be written in very simply matrix algebra as:

$$p(x) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right] \tag{5}$$

And this is the definition of the MVN!

In fact, eq. 5 is also the equation for a general MVN, even when the x_is are not independent, with one key difference: Σ becomes the covariance matrix. We will not explicitly derive this, just recall that the covariance matrix is defined as $\Sigma_{ij} = \mathrm{Cov}[x_i, x_j]$, that it is symmetrical and PD, meaning it is invertible. We can now change the notation again, so that it will become something slightly more compact:

$$p(x) = \mathcal{N}(x;\ \mu, \Sigma) \Leftrightarrow x \sim \mathcal{N}(\mu, \Sigma)$$

And here we really see why we constrained ourselves to PD covariance matrices - if it was not, the expression in eq. 5 would be undefined as the inverse of Σ would not be defined if it were not PD but only PSD (i.e., there exists an eigenvalue that equals 0).

Just to drive the point home, of course for any $x \sim \mathcal{N}(\mu, \Sigma)$ the following holds:

$$\mathbb{E}[x] = \mu \qquad \mathrm{Cov}[x] = \Sigma$$

and if is symmetrical. Since the PDF is once again symmetrical, we can try and think about the shape that would be the result of choosing all of the points with a probability higher than some value, i.e. looking at the shape defined by $P(x \geq a)$ for some a. In the multivariate case this shape becomes an ellipsoid, and in 2D it is an ellipse. In the special case when

$$\Sigma = \sigma^2 I$$

in other words when all of the random variables are independent and have the same variance, this becomes a ball.

Mahalanobis Distance

In the above section we have derived the PDF of the MVN and stumbled across a distance metric that is useful to know. This distance metric is the Mahalanobis distance and is defined as:

$$D_M(x; y) = \sqrt{(x - \mathbb{E}[y])^T \mathrm{Cov}[y]^{-1} (x - \mathbb{E}[y])}$$

where y is some random variable. This is a distance metric that tries to check how far the point x is from the distribution of y. In fact, the square of this distance is exactly how we will check what the likelihood of a point x is to be part of a Gaussian y, since it is what appears in the exponent of $p(y)$.

2.3 Sampling from the Gaussian Distribution

To show how to sample new points from an MVN, we will first look at an interesting property of the Gaussian distribution - a linear transformation of a Gaussian is, itself, a Gaussian.

Linear Transformations of the Gaussian Distribution

Given $x \sim \mathcal{N}(\mu, \Sigma)$, we want to know what the distribution of $y = Ax + b$ is, for any invertible A and any b. Recall from eq. 3 that if $y = f(x)$ and f is invertible, then we will be able to quantify the PDF of y in terms of the PDF of x. In this case, $f(x) = Ax + b$ so the inverse can be easily found:

$$\begin{aligned} y &= Ax + b \\ \Rightarrow y - b &= Ax \\ \Rightarrow A^{-1}(y - b) &= x \\ \Rightarrow f^{-1}(y) &= A^{-1}(y - b) \end{aligned}$$

We also need to check what the Jacobian of f^{-1} is[3]:

$$\frac{\partial}{\partial y} A^{-1}(y - b) = A^{-1}$$

Now we can find the distribution of y in the terms of the distribution of x:

$$\begin{aligned} p_y(y) &= \left|J\left(f^{-1}(y)\right)\right| p_x\left(f^{-1}(y)\right) \\ &= \left|J\left(A^{-1}(y - b)\right)\right| p_x\left(A^{-1}(y - b)\right) \\ &= \frac{1}{|A|}\mathcal{N}\left(A^{-1}(y - b);\quad \mu, \Sigma\right) \end{aligned}$$

Here I have used the fact that $\left|A^{-1}\right| = |A|^{-1}$. Okay, now we have to actually show that this PDF becomes a Gaussian distribution. Let's only look at the term that will be in the exponent for a moment:

$$\left(A^{-1}(y - b) - \mu\right)^T \Sigma^{-1}\left(A^{-1}(y - b) - \mu\right) = (y - b - A\mu)^T A^{-1T}\Sigma^{-1}A^{-1}(y - b - A\mu)$$

Let's denote $b + A\mu \equiv \eta$. At the same time, notice that since $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$, then $A^{-1T}\Sigma^{-1}A^{-1} = \left(A\Sigma A^T\right)^{-1}$, so the term in the exponent becomes:

$$= (y - \eta)^T\left(A\Sigma A^T\right)^{-1}(y - \eta)$$

Suddenly this seems very close to our original Gaussian distribution, right? Since this is the term in the exponent, we know have to show that the covariance in the normalizing term is also $A\Sigma A^T$:

$$\begin{aligned} \frac{1}{|A|} \cdot \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} &= \frac{1}{|A|^{1/2}\left|A^T\right|^{1/2}} \cdot \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \\ &= \frac{1}{\sqrt{(2\pi)^n |A| |\Sigma| \left|A^T\right|}} \\ &= \frac{1}{\sqrt{(2\pi)^n \left|A\Sigma A^T\right|}} \end{aligned}$$

In these equations have used the identities of $|A| = \left|A^T\right|$, as well as the fact that $|A|$ is a scalar, so I can move it about freely. Finally, I used the identity $|AB| = |A|\,|B|$.

Let's write everything together:

$$p_y(y) = \frac{1}{\sqrt{(2\pi)^n \left|A\Sigma A^T\right|}} \exp\left[-\frac{1}{2}(y - \eta)^T\left(A\Sigma A^T\right)^{-1}(y - \eta)\right]$$

This is very clearly a Gaussian distribution, and in fact we can now write $y \sim \mathcal{N}\left(\eta, A\Sigma A^T\right) = \mathcal{N}\left(A\mu + b, A\Sigma A^T\right)$.

[3] You should check this yourself. Finding the derivatives of matrices is, in general, troublesome. However, many of them are recorded in the wikipedia page Matrix Calculus or can be found in the Matrix Cookbook.