Propagation of uncertainty/compositions of random variables

August 21, 2018

Contents

1 Introduction and definitions

Stochastic methods often generate samples of random variables belonging to various distributions, call them x_i, where x_i is a vector of results from sample i. By averaging these samples, we generate a new random variable, $\overline{x}$. If the central limit theorem applies, we know that

$$\overline{x} \sim N(\mu, \Sigma); \qquad P(\overline{x}|\mu, \Sigma) = \frac{\exp\left\{-\frac{1}{2}(\overline{x}-\mu)^T\Sigma^{-1}(\overline{x}-\mu)\right\}}{(2\pi)^{d/2}|\Sigma|^{1/2}}.$$

μ and Σ represent the true mean and covariance matrix of the distribution that generated the samples, x_i.

Now we want to do calculations with these variables, e.g. compute a quantity $f(\overline{x})$. This new quantity will be a new random variable, and it is often useful to estimate the variance of this new distribution, and estimate how it fluctuates around $f(\mu)$. This is what I mean by **propagation of uncertainty**.

2 Linear functions

Let's represent the function as $f(\overline{x}) = \sum_j c_j \overline{x}^{(j)} = c \cdot \overline{x}$. Because expectations are linear,

$$\langle f(\overline{x})\rangle = c \cdot \langle \overline{x}\rangle = c \cdot \mu = f(\mu).$$

So the mean of the new distribution the same as if we had used μ. That's reassuring! But the variance about this value is also important:

$$\begin{aligned}
\langle (f(\overline{x}) - f(\mu))^2\rangle = \langle c\cdot(\overline{x}-\mu)^2\rangle &= \left\langle \left(\frac{1}{N}\sum_i c\cdot x_i - c\cdot\mu\right)^2\right\rangle \\
&= \frac{1}{N^2}\left\langle \sum_{ik}\left(c_k x_i^{(k)} - c_k\mu^{(k)}\right)^2\right\rangle \\
&= \frac{1}{N^2}\sum_{ijkl} c_k c_l \left\langle \left(x_i^{(k)} - \mu^{(k)}\right)\left(x_j^{(l)} - \mu^{(l)}\right)\right\rangle . \qquad (1)
\end{aligned}$$

Now assuming the samples x_i are independent and i.i.d., which is already required for the central limit theorem, this reduces to covariance matrix (Σ) of the variables $x^{(k)}$ divided by N, the number of samples when taking the mean. This is because the covariance of random variances in two samples is zero (independent), while the covariance between each sample is the same (knocking out factor one N).

$$\mathrm{Var}[f(\overline{x})] = \frac{1}{N}\sum_{kl} c_k c_l \Sigma_{kl} = \frac{1}{N} c^T \Sigma c. \quad \square$$

The variance of the final result is a weighted sum of the covariances of the arguments. The weights are the coefficients of the two things that are varied. A common case is that $f(\overline{x}) \propto \sum_k \overline{x}^{(k)}$. In that case, if two variables are perfectly negatively correlated, their errors will cancel!

Final note: you may notice this is different from by a factor of N from the wikipedia equation. This is because I've assumed you did some averaging before plugging things in. If you use the case $N = 1$ it recovers the case of $\mathrm{Var}[f(x)]$, where you apply f on each sample during your run.

2.1 Linearization

Nonlinear functions are sometimes approximated as linear functions using the Taylor expansion. A convenient place to expand is around μ because then things will cancel in the covariance. Later it will become apparent that this approximation is good for distributions that are peaked around μ, which is often the case.

$$f(\overline{x}) \approx f(\mu) + \sum_k \partial_k f(\mu)(\overline{x}^{(k)} - \mu) \tag{2}$$

A few steps will show you that in this case, $c_k = \partial_k f(\mu)$, and out pops the normal propagation of error formula (on wikipedia):

$$\Sigma^f = J \Sigma J^T$$

where Σ^f is the covariance matrix for f and $J_{ij} = \partial f^{(i)}/\partial x^{(j)}$ is the Jacobian of $f(x)$. In the case that Σ is diagonal, the even more common formula makes its appearance:

$$\mathrm{Var}[f(\overline{x})] = \sum_k (\partial_k f(\mu))^2 \mathrm{Var}[x^{(k)}].$$

$$\iff \sigma_f^2 = \sum_k \left(\frac{\partial f}{\partial x^{(k)}}\right)^2 \sigma^2_{x^{(k)}}$$

Of course, we often don't have $\partial_k f(\mu)$, so this is approximated as $\partial_k f(\overline{x})$. The accuracy of this approximation can be understood in the discussion of (3).

3 Taylor expansion

If your function is nonlinear but is sharply peaked around the mean, a Taylor expansion may be a good approximation. Taylor expansions are usually approximations that depend on taking powers of a small parameter. In this case the small parameter will be the various products of $(\overline{x}^{(i)} - \mu^{(i)})$, which will be close to zero, particularly when the average is over many samples. These equations are also more accurate the fewer nonzero derivatives f has. For example, $f(\overline{x}) = \overline{x}^2$ will have an exact formula that isn't too complicated. The first step is linearization[2.1], but that analysis doesn't cover errors or systematic improvement.

The multidimensional Taylor expansion takes the form:

$$f(x) = f(\mu) + \sum_{n_1,\ldots,n_d=1}^{\infty} \prod_i \frac{(x^{(i)} - \mu^{(i)})^{n_i}}{n_i!} \partial_i^{n_i} f(\mu)$$

The linearized form (2) comes about by dropping terms with more than one factor of $(x_i - \mu_i)$, which means (1) only first-order derivatives are involved, and (2) only linear terms are present. The resulting $\mathrm{Var}[f]$ only involved covariances because it arises from squaring a linear function. If we keep additional terms, therefore, the expressions will have higher moments of the distribution. These higher moments represent the errors of the linearization, and quickly go to zero as the distribution becomes sharply peaked around μ (which happens with the central limit theorem, for instance).

3.1 Corrections to linearization

To derive more accurate formula, choose where to truncate and follow a similar procedure to linear functions (Sec. 2). We'll need the third order term in order to get a correction to the variance of $f(\overline{x})$, since that is already second order.

$$\begin{aligned}\langle f(\overline{x})\rangle &\approx f(\mu) + \sum_i (\langle \overline{x}^{(i)}\rangle - \mu^{(i)})\partial_i f(\mu) + \sum_{i,j}\langle(\overline{x}^{(i)} - \mu^{(i)})(\overline{x}^{(j)} - \mu^{(j)})\rangle\partial_{ij} f(\mu) \\ &+ \sum_{i,j,k}\langle(\overline{x}^{(i)} - \mu^{(i)})(\overline{x}^{(j)} - \mu^{(j)})(\overline{x}^{(k)} - \mu^{(k)})\rangle\partial_{ijk} f(\mu)\end{aligned}$$

Like before, the first order is zero, while the third term becomes the same expression worked on in (1). Thus the second-order correction is related to covariance of the variables:

$$\begin{aligned}\langle f(\overline{x})\rangle &\approx f(\mu) + \sum_{i,j}\mathrm{Cov}[\overline{x}^{(i)}, \overline{x}^{(j)}]\partial_{ij} f(\mu) = f(\mu) + \frac{1}{N}\sum_{i,j}\mathrm{Cov}[x^{(i)}, x^{(j)}]\partial_{ij} f(\mu) \\ &= f(\mu) + \frac{1}{N}\mathrm{Tr}(\Sigma H), \quad \square \end{aligned} \tag{3}$$

where H is the Hessian of f.

To compute the next order correction to the variance, I square the expressions and drop everything of order higher than 3.

$$\begin{aligned}\langle f(\overline{x})^2\rangle &= f(\mu)^2 + \sum_{ij}\langle(\overline{x}^{(i)} - \mu^{(i)})(\overline{x}^{(j)} - \mu^{(j)})\rangle\partial_i f(\mu)\partial_j f(\mu) \\ &+ 2f(\mu)\sum_i\langle\overline{x}^{(i)} - \mu^{(i)}\rangle\partial_i f(\mu) \\ &+ 2f(\mu)\sum_{ij}\langle(\overline{x}^{(i)} - \mu^{(i)})(\overline{x}^{(j)} - \mu^{(j)})\rangle\partial_{ij} f(\mu) \\ &+ 2\sum_{ijk}\langle(x^{(i)} - \mu^{(i)})(\overline{x}^{(j)} - \mu^{(j)})(\overline{x}^{(k)} - \mu^{(k)})\rangle\partial_i f(\mu)\partial_{jk} f(\mu)\end{aligned}$$

The second-order term was already derived in Sec. 2. Computing $\langle f(\overline{x})^2\rangle - \langle f(\overline{x})^2\rangle$ will cancel everything multiplying $f(\mu)$. The first order term is zero as should be familiar by now. What's left over is

$$\mathrm{Var}[f(\overline{x})] = J\Sigma J^T + 2\sum_{ijk}\langle(x^{(i)} - \mu^{(i)})(\overline{x}^{(j)} - \mu^{(j)})(\overline{x}^{(k)} - \mu^{(k)})\rangle(\partial_i f(\mu)\partial_{jk} f(\mu) - f(\mu)\partial_{ijk} f(\mu)). \quad \square$$

Thus the variance formula is also only exact for linear functions, and the very next order is nonzero.

4 Bayesian error analysis

5 Bootstrapping

6 Lemmas

Here's some generic results that are useful but not important to understand the general idea.

7 Expectations of products of independent random variables

Being independent means that for a set of random variables $x = \{x_i\}$, $p_X(x) = \prod_i p_{X_i}(x_i)$, i.e. the probability distribution factorizes. As we'll see this implies that $\mathrm{Cov}[x_i, x_j] \propto \delta_{ij}$, but the inverse is not necessarily true because it also says something about all the higher moments. Without loss of generality, let x_i all have mean zero, because otherwise we can define a new variable $y_i = x_i - \mu_i$ to shift it to zero.

Being independent,

$$\left\langle\prod_i x_i^{n_i}\right\rangle = \int \rho_X(x)\prod_i dx_i x_i^{n_i} = \prod_i\int dx_i\rho_{X_i}(x_i)x_i^{n_i} = \prod_i\langle x_i^{n_i}\rangle. \quad \square$$

This implies $\text{Cov}[x_i, x_j] = 0$ because the distributions each have mean zero. But the inverse is not true because among other things, all moments involving $n_i = 1$ for any i will also be zero.