Tyler Stanish
MA 351

Applying Bayesian Inference to Parametric Models: An Example with Least Squares Regression

Most machine learning and statistical methods use a dataset to perform computations that eventually lead to a model. With the exception of "online learning," models are performed "at once" and are then used to predict events given new data. In many industrial settings where there is a vast amount of data and new incoming data continuously, many models do not take into consideration new data as it arrives. Rather, these models must collect the new batch of data, parse the data, and feed it into a database where it is queried and retrained. This paper attempts to provide a barebones approach to a Bayesian method of updating already trained models without having to retrain.

We take as an example the least squares regression model. This model is easy to compute mathematically but can of course be abstracted to other models given that these models can be parametrized. This method is not recommended for linear regression because the parameters can be calculated "on the fly" quite easily without Bayesian inference.

To represent some linear function of the data we take the model:

$$\boldsymbol{y} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \beta_{n+1} + \varepsilon$$

$\boldsymbol{\beta}$ is the vector of true parameters: $\boldsymbol{\beta} = [\beta_1 \quad \beta_2 \quad \ldots \quad \beta_{n+1}]^T$. ε is the error term; it is normally distributed according to $\mathcal{N}(0, \sigma^2)$. Because we are trying to estimate the true parameters, $\hat{\boldsymbol{\beta}}$ will be used for our estimation of the vector of parameters. Now we can represent our least squares model in vector/matrix form:

$$X\hat{\boldsymbol{\beta}} = \boldsymbol{y}$$

$$X = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{m2} & \cdots & X_{mn} \end{pmatrix}, \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_{n+1} \end{pmatrix}, \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

We solve for the vector of parameters using the normal equations:

$$\hat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T \mathbf{y}$$

Because each element of $\hat{\boldsymbol{\beta}}$ is a random variable (all estimators are random variables because all samples are taken randomly), $\hat{\boldsymbol{\beta}}$ is a random vector. Just as we might assume the estimator of some univariate model to be normally distributed (given large enough data), we can assume the estimator of some multivariate model to be normally distributed. The multivariate analog of the normal distribution is as follows[1]:

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

with p being the number of predictors; in our case it is $n+1 = \dim(\hat{\boldsymbol{\beta}})$. $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ are the parameters of the multivariate normal distribution. $\boldsymbol{\mu}$ is the population mean vector, and $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ is the variance-covariance matrix; it is a positive definite matrix which can be constructed for the estimation vector from the least-squares fitting by[2]:

$$\boldsymbol{\Sigma} = \sigma^2 \left(X^T X\right)^{-1}$$

The matrix $(X^T X)^{-1}$ is also known as the inverse of the Gramian matrix[3] of X. We also take the other parameter of the multivariate Gaussian distribution, $\boldsymbol{\mu}$, to be the values calculated from $\hat{\boldsymbol{\beta}}$.

Now suppose that after training our model we receive new data with which we would like to update our parameter vector. With this new set of data (let's denote it as $\boldsymbol{x}$), we can repeat this process but fitting the new parameters only from the new data. This is demonstrated in the same fashion as the initial training shown above:

$$\hat{\boldsymbol{\beta}}_{\mathbf{new}} = \left({X_{\text{new}}}^T X_{\text{new}}\right)^{-1} {X_{\text{new}}}^T \mathbf{y}_{\mathbf{new}}$$

Let the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathbf{new}}$ be $\boldsymbol{\Omega}$; it is also derived in the same fashion as above. Now is where Bayesian inference comes in; here is the general formula of computing the posterior distribution[4]:

$$p\left(\hat{\boldsymbol{\beta}}_{\mathbf{0}} \mid \{\boldsymbol{x}_i\}\right) \propto p\left(\hat{\boldsymbol{\beta}}_{\mathbf{0}}\right) \prod_{i=1}^{N} p\left(\boldsymbol{x}_i \mid \hat{\boldsymbol{\beta}}_{\mathbf{0}}\right)$$

Because we are training the new batch of data only once, $N=1$. Also, instead of $\boldsymbol{x}$ we have $\hat{\boldsymbol{\beta}}_{\mathbf{new}}$ and instead of $\hat{\boldsymbol{\beta}}$ we have $\hat{\boldsymbol{\beta}}_{\mathbf{0}}$ which will be used from now on to refer to the parameters of the prior multivariate Gaussian (determined from the first training of the model, because it is our *prior* belief before the new data arrives). Therefore, our formula reduces to:

$$p\left(\hat{\boldsymbol{\beta}}_{\mathbf{0}} \mid \hat{\boldsymbol{\beta}}_{\mathbf{new}}\right) \propto p\left(\hat{\boldsymbol{\beta}}_{\mathbf{0}}\right) p\left(\hat{\boldsymbol{\beta}}_{\mathbf{new}} \mid \hat{\boldsymbol{\beta}}_{\mathbf{0}}\right)$$

The product of a multivariate Gaussian prior and a multivariate Gaussian likelihood function yields a multivariate Gaussian distribution as the posterior (in more formal terms, the Normal distribution is its own *conjugate prior*). The values that we plug back into the model are those that are determined *a posteriori*, a simple method known as maximum a posteriori (MAP) estimation:

$$\hat{\boldsymbol{\beta}}_{\text{MAP}} = \operatorname{argmax}_{\hat{\boldsymbol{\beta}}}\, p\left(\hat{\boldsymbol{\beta}}_{\mathbf{0}} \mid \hat{\boldsymbol{\beta}}_{\mathbf{new}}\right)$$

In general, this is the computational formula of how we update a multivariate normal distribution with a conjugate prior of a multivariate normal distribution[5]:

$$\widehat{\boldsymbol{\beta}_n} = \boldsymbol{\Sigma}\left(\boldsymbol{\Sigma} + \frac{1}{N}\boldsymbol{\Omega}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\beta}}_{new\,i}\right) + \frac{1}{N}\boldsymbol{\Omega}\left(\boldsymbol{\Sigma} + \frac{1}{N}\boldsymbol{\Omega}\right)^{-1}\hat{\boldsymbol{\beta}}_{\mathbf{0}}$$

$$\boldsymbol{\Sigma}_n = \boldsymbol{\Omega}\left(\boldsymbol{\Sigma} + \frac{1}{N}\boldsymbol{\Omega}\right)^{-1}\frac{1}{N}\boldsymbol{\Sigma}$$

which can be derived using matrix algebra[6] (where the $\boldsymbol{\mu}$ of our distribution is $\hat{\boldsymbol{\beta}}$):

$$\begin{aligned}\ln p\left(\hat{\boldsymbol{\beta}} \mid \hat{\boldsymbol{\beta}}_{\text{new}}\right) &\propto -\frac{1}{2}\sum_{i=1}^{N}\left(\hat{\boldsymbol{\beta}}_{\text{new}i} - \hat{\boldsymbol{\beta}}\right)^T\boldsymbol{\Omega}^{-1}\left(\hat{\boldsymbol{\beta}}_{\text{new}i} - \hat{\boldsymbol{\beta}}\right) - \frac{1}{2}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathbf{0}}\right)^T\boldsymbol{\Sigma}^{-1}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathbf{0}}\right)\\ &\propto -\frac{1}{2}N\hat{\boldsymbol{\beta}}^T\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\beta}} + \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}\hat{\boldsymbol{\beta}}^T\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\beta}}_{\text{new}i} - \frac{1}{2}\hat{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}_{\mathbf{0}}\\ &\propto -\frac{1}{2}\hat{\boldsymbol{\beta}}^T\left(N\boldsymbol{\Omega}^{-1} + \boldsymbol{\Sigma}^{-1}\right)\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^T\left(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}_{\mathbf{0}} + \boldsymbol{\Omega}^{-1}\sum_{\boldsymbol{i=1}}^{N}\hat{\boldsymbol{\beta}}_{\text{new}i}\right)\end{aligned}$$

Let $a = \left(N\boldsymbol{\Omega}^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}$ and let $b = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}_{\mathbf{0}} + \boldsymbol{\Omega}^{-1}\sum_{\boldsymbol{i=1}}^{N}\hat{\boldsymbol{\beta}}_{\text{new}i}$

$$\propto -\frac{1}{2}\left(\hat{\boldsymbol{\beta}} - ab\right)^T a\left(\hat{\boldsymbol{\beta}} - ab\right)$$

which we recall from above (you may notice $\hat{\boldsymbol{\beta}}$ is on the left side of the minus sign, this is because we are finding $\ln p\left(\hat{\boldsymbol{\beta}} \mid \hat{\boldsymbol{\beta}}_{\text{new}}\right)$ rather than $p_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x})$), this is the log of the multivariate normal:

$$p\left(\hat{\boldsymbol{\beta}} \mid \hat{\boldsymbol{\beta}}_{\text{new}}\right) \sim \mathcal{N}(ab, a)$$

Now using the Woodbury identity for a, the covariance matrix:

$$a = \left(N\boldsymbol{\Omega}^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1} = \boldsymbol{\Omega}\left(\frac{1}{N}\boldsymbol{\Omega} + \boldsymbol{\Sigma}\right)^{-1}\frac{1}{N}\boldsymbol{\Sigma}$$

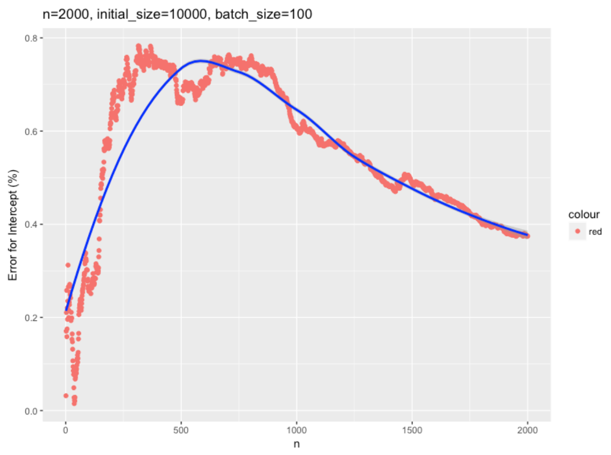
which is in the same form as shown above ($\boldsymbol{\Sigma}_n$). And now the mean:

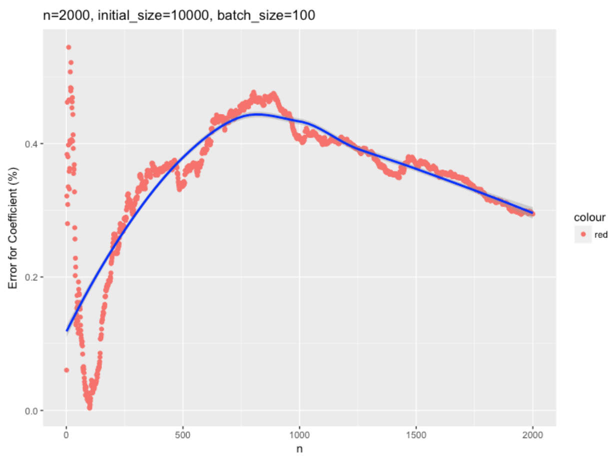
$$\begin{aligned}a &= \boldsymbol{\Omega}\left(\frac{1}{N}\boldsymbol{\Omega} + \boldsymbol{\Sigma}\right)^{-1}\frac{1}{N}\boldsymbol{\Sigma}\\ ab &= \boldsymbol{\Omega}\left(\frac{1}{N}\boldsymbol{\Omega} + \boldsymbol{\Sigma}\right)^{-1}\frac{1}{N}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}_{\mathbf{0}} + \boldsymbol{\Sigma}\left(\frac{1}{N}\boldsymbol{\Omega} + \boldsymbol{\Sigma}\right)^{-1}\frac{1}{N}\boldsymbol{\Omega}\boldsymbol{\Omega}^{-1}\sum_{\boldsymbol{i=1}}^{N}\hat{\boldsymbol{\beta}}_{\text{new}i}\\ &= \boldsymbol{\Sigma}\left(\boldsymbol{\Sigma} + \frac{1}{N}\boldsymbol{\Omega}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\beta}}_{\text{new}i}\right) + \frac{1}{N}\boldsymbol{\Omega}\left(\boldsymbol{\Sigma} + \frac{1}{N}\boldsymbol{\Omega}\right)^{-1}\hat{\boldsymbol{\beta}}_{\mathbf{0}}\end{aligned}$$

also in the same form as above ($\widehat{\boldsymbol{\beta}_n}$). In our case, because we are not using the data but the trained parameters of a least squares model over the new data (not including the old data) again $N = 1$ because we only have one $\hat{\boldsymbol{\beta}}_{\text{new}}$. I include N because I want to allow the reader the freedom to partition the data into multiple training sets.

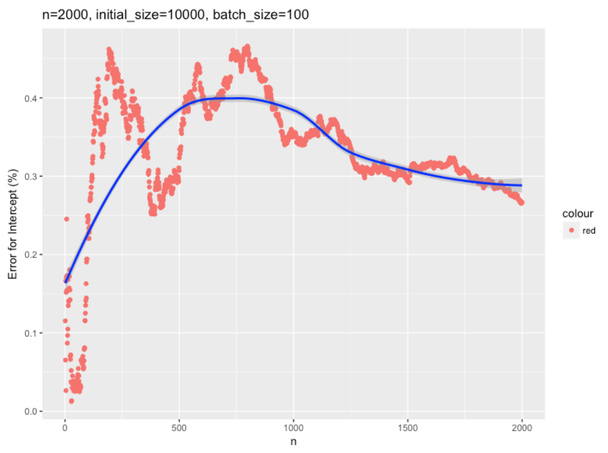
Results

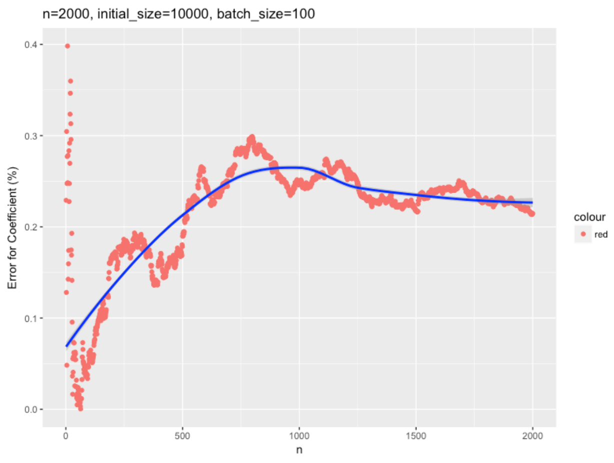
For 2000 repeated trials (where a trial consists of calculating the posterior distribution and "plugging in" the $\hat{\boldsymbol{\beta}}_{\mathrm{MAP}}$ values into the new model), an initial training sample size of 10,000 and a new batch size of 100 (the square root of 10,000, because it is likely that the amount of data collected over time follows a square root law) the results were:



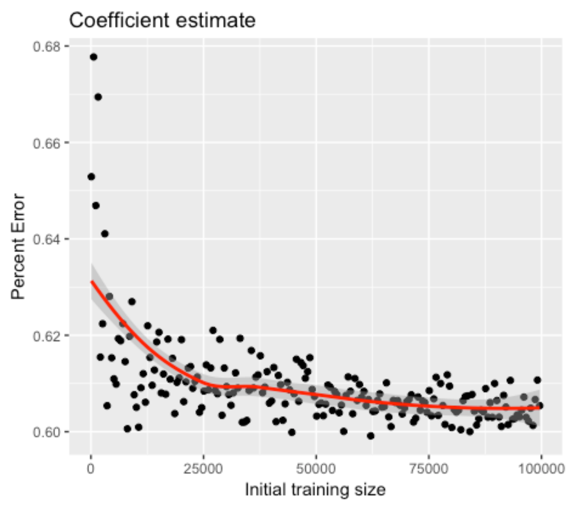


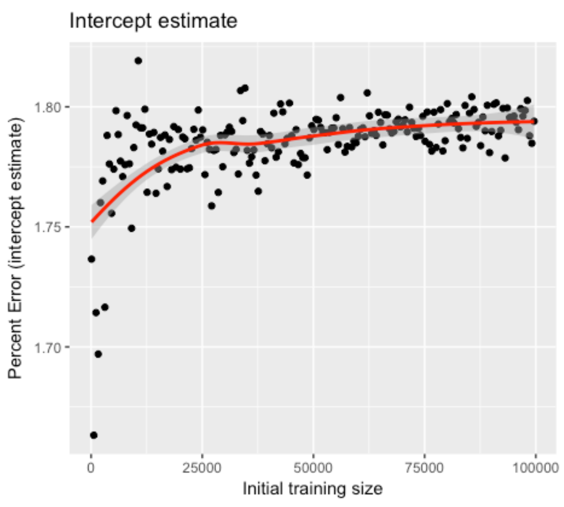
You will notice that these graphs seem to be *processes* or *random walks* rather than random scatter plots, and this is because if we introduce error to a single posterior distribution, it is then fed back in as the prior of the next iteration which brings the bias from the last posterior which may in turn re-introduce more (or less) error. This is perhaps the reason why the shapes of these curves can change dramatically from one simulation to the next:





Also holding n=1 constant, we can graph percent error against initial training size with each new batch size being the square root of each individual initial training size:





The R code for generating these plots along with other simulation logic can be found at https://github.com/tylerstanish/bayesian-updating-MA351-project.git

Notes

It is important to note that even though there is a small percent error, there was also a small increase in the data provided and thus a small change in the parameters from the "true" trained models.

Citations

1. James, G., Witten, D., Hastie, T., Tibshirani, R.: An introduction to statistical learning, vol. 112. Springer, New York (2013)
2. Del Pino, Guido E., and Hector Galaz. "STATISTICAL APPLICATIONS OF THE INVERSE GRAM MATRIX: A REVISITATION." *Brazilian Journal of Probability and Statistics*, vol. 9, no. 2, 1995, pp. 177–196. *JSTOR*, JSTOR, www.jstor.org/stable/43601459.
3. Ibid.
4. Marin, Jean-Michel and Robert, Christian P.: Bayesian Core, first edition. Springer, New York (2007).
5. Alex (https://stats.stackexchange.com/users/11408/alex), Multivariate normal posterior, URL (version: 2017-03-21): https://stats.stackexchange.com/q/28744
6. conjectures (https://stats.stackexchange.com/users/16663/conjectures), Multivariate normal posterior, URL (version: 2017-03-21): https://stats.stackexchange.com/q/268843